

University of Pretoria etd – van der Merwe, A (2008)

**DEVELOPMENT AND IMPLEMENTATION OF AN INSTITUTIONAL
REPOSITORY WITHIN A SCIENCE, ENGINEERING AND TECHNOLOGY
ENVIRONMENT**

Mini-dissertation by

**Adèle van der Merwe
(7622449)**

Submitted in partial fulfilment of the requirements for the degree

MAGISTER PHILOSOPHIAE (INFORMATICS)

in the

DEPARTMENT OF INFORMATICS

of the

**FACULTY OF ENGINEERING, BUILT ENVIRONMENT
AND INFORMATION TECHNOLOGY
UNIVERSITY OF PRETORIA**

Supervisor: Prof. J.H. Kroeze

April 2008

Acknowledgements

Special words of gratitude go to:

- My sons Barend and Renier van der Merwe: Thank you for your support and willingness to settle for fast foods.
- Family and friends: Your belief in my abilities, as well as your support over the last three years while I was studying, meant a lot to me.
- Martie van Deventer, Madelein van Heerden, Yvonne Halland, Annette Joubert and Ina Smith: Without your advice, contributions and brain-storming the IR project would not have been the success it was.
- Siphethile Muswelanto: Thank you for standing in for me during an unforeseen crisis and for your initiative and assistance. Having somebody like you as an assistant was not only a pleasure but also a priceless advantage.
- Ciaran Mac Carron: Thank you for your strict but kind editing of my work.

TABLE OF CONTENTS

| | |
|---|-----------|
| GLOSSARY | v |
| 1 INTRODUCTION AND PROBLEM STATEMENT | 1 |
| 1.1 Problem Statement | 1 |
| 1.2 Objectives..... | 2 |
| 1.3 Outline of Scope..... | 3 |
| 1.4 Limitations of this Research | 5 |
| 1.5 Theories | 6 |
| 1.6 Research Approach and Research Questions..... | 7 |
| 1.7 Methodological Approach..... | 8 |
| 1.8 Design and Layout | 10 |
| 1.9 Conclusion..... | 10 |
| 2 LITERATURE REVIEW | 12 |
| 2.1 Relevant Literature..... | 12 |
| 2.1.1 Introduction to the relevancy of institutional repositories | 12 |
| 2.1.1.1 Defining the concept ‘Institutional Repository’ | 12 |
| 2.1.1.2 The emergence of Institutional Repositories..... | 14 |
| 2.1.1.3 Benefits and value of an IR | 16 |
| 2.1.1.4 Generic features of institutional repositories | 17 |
| 2.1.1.5 Stakeholders | 18 |
| 2.1.2 Development and planning of an IR..... | 18 |
| 2.1.3 Development worksheet | 20 |
| 2.1.4 Challenges facing implementation of an IR | 22 |
| 2.1.5 Technological issues..... | 25 |
| 2.1.5.1 Proprietary vs. Open Source Software..... | 26 |
| 2.1.5.2 IR software | 28 |
| 2.1.5.3 DSpace..... | 31 |
| 2.1.6 Policy issues | 33 |
| 2.1.6.1 Essential elements in policies | 34 |
| 2.1.7 Cost models | 36 |
| 2.1.8 Preservation issues | 37 |
| 2.1.9 Compliance and content recruitment issues | 40 |
| 2.1.9.1 Pre-publications and other issues regarding copyright materials | 42 |
| 2.1.10 Operational issues..... | 43 |
| 2.1.10.1 Quality control and standardization..... | 43 |
| 2.1.10.2 Auditing | 45 |
| 2.2 Identification of the Gap between Problem Statement and Literature | 47 |
| 2.3 Conclusion..... | 49 |
| 3 DEVELOPMENT OF A REPOSITORY..... | 50 |
| 3.1 Introduction..... | 50 |
| 3.2 User Requirements and Specifications (URS)..... | 53 |
| 3.2.1 Structure and features of repository | 55 |
| 3.2.2 Internal controls | 57 |
| 3.3 Policies, Procedures and Managerial Reports..... | 59 |
| 3.4 Compliance | 61 |
| 3.4.1 Procedures and processes | 62 |
| 3.4.2 Standards and quality control of metadata | 67 |
| 3.5 Determining success..... | 68 |
| 3.5.1 Visibility and Usage Statistics..... | 69 |
| 3.6 Skills development | 74 |
| 3.7 Conclusion..... | 75 |

| | | |
|-----|--|-----|
| 4 | RESULTS AND DISCUSSION | 77 |
| 4.1 | Lessons learned..... | 78 |
| 4.2 | Problems experienced | 79 |
| 4.3 | Evaluation and reasoning..... | 80 |
| 4.4 | Recommendations | 81 |
| 4.5 | Expected Contribution to Knowledge | 84 |
| 4.6 | Future research..... | 85 |
| | 4.6.1 Determining the h-index..... | 85 |
| | 4.6.2 Determining the Return-on-Investment..... | 85 |
| 4.7 | Conclusion..... | 86 |
| 5 | CONCLUSION | 87 |
| 6 | REFERENCES..... | 91 |
| | Attachment A: CSIR Research Space Policy: Stakeholders and compliance | 100 |
| | Attachment B: CSIR Research Space Policy: Metadata, data, submission and preservation..... | 106 |
| | Attachment C: Structure of CSIR Research Space..... | 108 |
| | Attachment D: Complete list and usage by country - November 2007 | 110 |

List of figures

| | |
|---|----|
| Figure 1: Content types as reflected in DOAR repositories | 25 |
| Figure 2: Graphical representation of Repository Software | 28 |
| Figure 3: DSpace Ingest process | 32 |
| Figure 4: Submission workflow in DSpace | 33 |
| Figure 5: Research Space homepage | 51 |
| Figure 6: CSIR Homepage | 51 |
| Figure 7: CSIR Research Space structure | 55 |
| Figure 8: Authorization management by means of policies | 58 |
| Figure 9: Access rights according to groups | 59 |
| Figure 10: TOdB/CSIR Research Space linked workflow, taken from the work document | 64 |
| Figure 11: CSIR Research Space workflow | 66 |
| Figure 12: Graphical representation of usage | 70 |
| Figure 13: Search engine connections - November 2007 | 72 |
| Figure 14: Daily visits - November 2007 | 73 |

List of tables

| | |
|--|----|
| Table 1: Comparison of functionalities: InMagic (proprietary) and DSpace (OSS) | 27 |
| Table 2: RoMEO Colours | 42 |
| Table 3: Dublin Core Metadata elements – Mandatory elements | 44 |
| Table 4: Dublin Core Metadata elements – Optional elements | 44 |
| Table 5: Authorization in DSpace | 46 |
| Table 6: Usage August 2007 - November 2007 | 69 |
| Table 7: Monthly breakdown for August-November 2007 | 70 |
| Table 8: International access - Status in November 2007 | 71 |
| Table 9: Items with the 12 highest view scores | 73 |

GLOSSARY

Abbreviations

BOAI: Budapest Open Access Initiative

CSIR: Council for Scientific and Industrial Research

CSIRIS: CSIR Information Services

DCMI: Dublin Core Metadata Initiative

EBAS: Enterprise-Based Applications and Systems

FOSS: Free Open-Source Software

ICT: Information, Communication, and Technology

IP number: Internet Protocol number

IPR: Intellectual Property rights

IR: Institutional Repository

IS: Information System

ISBN: International Standard Book Number

KPI: Key Performance Index

KRA: Key Result Areas

OA: Open Access

OpenDOAR: Directory of Open Access Repositories

OSI: Open Society Institute

OSS: Open Source Software

R&D: Research and Development

ROI: Return on Investment

SET: Science, Engineering and Technology

SHERPA: Securing a Hybrid Environment for Research Preservation and Access

TOdB: Technical Outputs Database

UP/AIS: University of Pretoria's Academic Information Services

URI: Unique Resource Identifier

URL: Uniform/Universal Resource Locator

URS: User Requirement Specifications

Working definitions

Bitstreams: In the context of this document, digital data transmitted between systems, irrespective of format.

Copyright owner: The person/institution holding the copyright, not necessarily the author/institution who owns the intellectual property right.

Curate/Curation: In this case identification, collection, safekeeping and management of explicit information sources, normally done by a curator.

End-users: The individuals/data consumers for whom the product was intended vs. intermediaries preparing/selling/managing the product.

e-Science: is used to describe scientific endeavours that requires distributed networks and/or that uses immense data sets

Explicit knowledge: Knowledge/information captured in print or other media.

Harvest: Retrieval and transfer of metadata from one database to another.

H-Index: The quantification of scientific productivity and impact of an individual scientist.

Hybrid system: The combination of a fee-based or managed accessed system and a free or open access system.

Ingest: The workflow process that accepts and loads a digital file.

Intellectual Property: The intangible creativity/knowledge of an individual.

Interested parties: Wider than peers only, as the man-on-the-street with a passing interest in specific information is included. See also end-users.

Just-in-case: Subscription to a journal because it might contain valuable information.

Just-in-time: Provision of relevant and specific information in a digital format on demand, usually within twenty-four hours.

Knowledge artefacts: Codified/explicit knowledge, irrespective of format

Metadata: Information describing a physical item, sometimes also called 'data about data'

Open URL Resolver: an actionable URL service based on metadata

Peer review: The reviewing process done by academic peers to determine the quality and accuracy of an article.

Post-print: The final accepted and published version of an article.

Pre-print: The final edited version of an article prior to publication and peer review

Preservation: In this scenario, ensuring that the digital format is accessible in future by means of migration or any other action deemed necessary.

Self-archiving: The process whereby authors can submit the metadata and full-text item of their own publications into a database

Shared understanding: Ensuring that all the parties understand exactly what is meant and that they do not attach their own interpretation to terminology.

SHERPA RoMeo: Publisher's copyright & archiving policies developed and maintained by SHERPA

Stakeholders: used as a collective noun to include data/information producers, including authors, data consumers and funding agencies

Summary

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 study provides a clear understanding of the issues surrounding the implementation of an IR. Issues discussed include software selection, as well as the development, implementation and marketing of an IR. An equally important issue is individual skills development. Attention is given to the development of the policies that are required by an organization and its main stakeholders. These policies form an essential part of the development of an information system. Issues such as acceptance, usage, population, and management of the repository are reported on. The actual work that was done at the CSIR is used as a case study. The implementation process at the CSIR 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 discussed.

1 INTRODUCTION AND PROBLEM STATEMENT

1.1 Problem Statement

'Doing what little one can to increase the general stock of knowledge is as respectable an object of life, as one can in any likelihood pursue.'

Charles Darwin (Darwin, n.d.)

The CSIR is a knowledge-based Science, Engineering, and Technology (SET) organization. One of the tangible outputs generated by the organization is explicit knowledge. The pressure on researchers to produce explicit knowledge artefacts contributes towards the plethora of explicit knowledge available in today's environment. The challenge facing most organizations, including the CSIR, is managing the vast volume of knowledge artefacts, ranging from research reports, books, research papers and conference papers to *ad hoc* explicit records in a variety of electronic formats, including e-mails, faxes and audiovisual material. Not only is effective management required but also the preservation and curation of the knowledge artefacts in order to facilitate efficient retrieval and to benefit from shared knowledge. Although preservation and curation have always been important, the preservation and curation of digital formats play an essential role in future research activities.

Researchers often express the need of the international scientific and research community to be able to monitor and share in the work that their peers are doing. This trend is evident in the emergence of the *h-index* developed by Hirsch (2005). In order to receive the recognition of their peers, organizations are expected - and perhaps even required - to be able and willing to: a) share their research and b) to provide the access required. The value of explicit knowledge artefacts is evident in all scientific fields. Value is determined by peer recognition and it is often regarded as the most important form of recognition. Positive and value-based recognition of the organization, the individual researcher or even of the research teams has become essential in a very competitive environment. In order to meet these demands, an Information System (IS) for the efficient and functional storage and retrieval system should be in place.

The emergence of Institutional Repositories (IR) is the result of an attempt to address some of these demands. However, the exact structure and function of such

a repository is subjected to individual organizational cultures and to their definition of a repository's purpose and functionality. It is therefore essential to identify and define all the relevant issues relating to the development of a repository and to determine the relevancy of these issues to an organization such as the CSIR. Furthermore, it is necessary to determine if an IR will fulfil the needs expressed above and to what extent.

1.2 Objectives

The primary objective of this study is therefore to investigate the current international trends for an institutional repository within a scientific, SET based, research organization. This will be done in terms of software and user requirements, developmental and implementation policies, and implementation and marketing issues. The focus is on the development of the IS (the IR) rather than on the skills generally associated with Information Science. It is important to note that, within the context of this dissertation, an IR cannot exist without a suitable information system being in place. However, out of necessity, reference will be made to some issues normally associated with the Information Science field. It is impossible to develop and implement an IR without an in-depth understanding of the underlying issues and factors that exist. Therefore, an extensive collaboration between Information Science, Information, Communication and Technology (ICT) and stakeholders are required, as IRs are just a small part of the general move towards E-science within a virtual environment.

Articles in scholarly journals show an increasing exaltation of the value and benefits of OSS products. Simultaneously an increase in the demand for open access to information and data is emerging. It is important that open access is not confused with free access, although the two concepts normally go hand-in-hand and are often regarded as synonyms. The Budapest Open Access Initiative (BOAI) refers to open access as the '...free and unrestricted online availability ...' of data and to the fact that authors will be given '...vast and measurable new visibility, readership and impact' (BOAI, 2007). The initiative is supported by two strategies: namely self-archiving, whereby authors are provided with the infrastructure to submit suitable material in an open-access archive or, alternatively, the right and freedom to publish their publications in accredited open-access journals. The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities expands on this approach by stating that contributions should 'include original scientific research

results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material' (Max Planck Society, 2003).

Secondary objectives of this study are:

- To identify and select the most suitable software and platform within the Open Source movement;
- To determine and develop the policies that will be required by the organization and its main stakeholders;
- To monitor skills development and career growth; and
- To identify the essential IS project management aspects required to address issues regarding the acceptance, use, population and management of the repository, e.g. change management.

1.3 Outline of Scope

First, it is essential to determine the technological demands that an IR must meet in order to be successful. With the Open Source Initiative (OSI) gaining momentum, one of the challenges facing organizations is the development of an institutional repository using open-source software and running on an open-source platform. The selected product should address and alleviate the shortcomings of earlier Open Source Information System products and should prove to be both cost effective and efficient. Some of the major challenges in selecting a suitable platform relate to compatibility issues, limitations of the existing infrastructure, availability of skilled resources and the fact that the open-source movement is still in its infancy. It is therefore essential to determine whether existing open-source repository products will meet the expectations and growing demands of IR champions. The platform selected will also have to satisfy criteria such as security, access control, adaptability, ease-of-use and long term archiving and accessibility.

The second challenge relates to the policies (or lack thereof) within the organization as applicable in the context of IS. Where policies are not consistently enforced or where there is an inadequate awareness of the existence of these, problems can arise, e.g. lack of standardization regarding formats and software, lack of compliance and lack of proper infrastructure support. It is therefore necessary to identify methods or tools to rectify this problem. The implementation of additional policies might also be required. Existing policies could require updating, because of

University of Pretoria etd – van der Merwe, A (2008)

unanticipated and unavoidable changes resulting from changes in technology, awareness, commitment and perceived needs. The scope of the policies adopted will be addressed. Some of the issues that are considered include the technological requirements, behavioural issues, stakeholder expectations and incentive statements of the organization. Although more of a peripheral nature, the absence of behaviour modification might influence the success of the project.

However, in order for developers and the operators to have the insight required, the organization's management team will have to be clear and specific regarding its requirements. To date the CSIR's IR project team has been an informal, *ad hoc* group of interested and enthusiastic individuals. It might therefore be necessary to form a more formal project team to achieve success with the implementation of the IR. In addition, it was essential that the team learned from identified and historical mistakes made during the development of existing and previous databases, especially when formalizing the policies and structure of the repository. Representatives of all the relevant stakeholders were involved in the project team on a needs basis.

The third challenge facing the CSIR's IR development team and the organization was that of ensuring the effective and efficient use of the repository, as well as delivery of the required outputs. Issues that were considered include methodologies to ensure quality, standardization and compliance. It is therefore important for the IS to manage these issues. To date, compliance in terms of contribution and submission to the bibliographical database has been a cause of concern. The lack of standards and compliance invariably results in the provision of incomplete or faulty information and records, which could result in a poor performance evaluation. It is therefore necessary to understand why there is a resistance on the side of authors, as indicated by their poor participation. In addition, it is necessary to understand why there is a general lack of adherence to policies and why information is not supplied in the required formats. This is partially done by determining whether the IR meets the expectations of the stakeholders and whether the correct platform was selected. Mutually agreed upon procedures must be identified in order to prevent a recurrence of existing negative behavioural patterns. It is necessary to determine whether the product is the reason for the poor behaviour or whether there are softer, behaviour-based problems.

A lack of visibility has potentially negative results in terms of the reputation and recognition of the organization. As with the implementation of any IS product, the champion of the repository will have to create a true appreciation and awareness of the benefits associated with compliance amongst the knowledge workers within the organization. Factors that make repositories effective include: a) clearly stated systems requirements, b) software that meets the demands and expectations of the participants and that can grow with changing demands, c) 'enthusiastic' compliance and support from the knowledge workers and the organization's management teams and d), the implementation of standards and policies that address expectations, incentives and recognition. A lack of interaction and cooperation between the team leader (as a representative of the developers) and the champion (as a representative of the stakeholders) would most likely lead to a breakdown in communication resulting in misunderstanding and misinterpretation of the needs and purpose of the repository. At best, this would result in a low return on investment but it could also result in outright rejection of the repository. A break in trust and support might be impossible to repair and could lead to the ultimate failure of the project.

1.4 Limitations of this Research

The development of an IR is a work-in-progress. It is impossible to reach a point at which complete satisfaction can be announced and where development of the repository can be regarded as 100% complete and where pure maintenance kicks in. Continuously changing needs, constant improvements in software and increasing demands by end-users, management, and authors place an ever-present demand on adapting the functionalities and features of the repository. It was therefore necessary to restrict this particular study to a fixed time period.

It is extremely difficult to determine the Return-on-Investment (ROI) in terms of an IR, as the value knowledge and information is difficult to measure. It is also difficult to measure the impact that individual authors are making within their respective research areas and the potential impact of an IR in this area. There is a plethora of information regarding the challenges facing long-term preservation. However, there is a lack of affordable solutions. Unless this problem is solved in an affordable and efficient manner, any repository is doomed to eventual failure. Additional systems development is required in all three cases. These issues will be discussed in more detail in the section on future research.

1.5 Theories

An overview of the concept of an institutional repository, together with an understanding of the technological issues involved, placed in context within certain assumptions made on a daily basis, is required for a true understanding of institutional repositories. Institutional repositories and the open-source initiative are regarded as two sides of the same coin. Ideally, both issues should be investigated simultaneously. Unfortunately, as this was not feasible, certain assumptions regarding prior knowledge of the open-source initiative had to be made. Clarity regarding the features, structure, principles and policies associated with institutional repositories were available in the form of empirical data and case studies. Unstructured interviews were used to obtain additional data and insight regarding anticipated use and value, as well as of obstacles in the acceptance of a repository.

Empirical data will mainly be used to obtain information regarding the existing developmental status of relevant software. Software specifications will enable the different products to be analysed and compared and will facilitate the selection and recommendation of the most suitable product. The empirical data that will be used are aspects such as proven compliance with operating systems such as Linux, hardware requirements, compatibility with the existing infrastructures and available support.

Available case studies of some of the bigger institutions and organizations will be analysed in order to become familiar with current approaches, problem areas identified and potential workable solutions to be proposed. Published reports by institutions that have already implemented a repository can potentially be used to identify applicable bottlenecks within the organization.

Literature surveys will be done on full text, peer-reviewed e-journal platforms such as ScienceDirect and ISI Web of Knowledge. Scientifically based Internet search engines, e.g. Scopus and Scirus, will be used to obtain the case studies and empirical data required, based on a tested literature search strategy. By making use of the alerting services of the above-mentioned resources, relevant and up-to-date information will be available for use. Because of the existing dynamic development phase surrounding repositories, up-to-date information is an essential tool and method for staying abreast of new developments, directions and implementation.

1.6 Research Approach and Research Questions

In order to obtain an acceptable level of understanding and insight into the complexities of an institutional repository, a critical interpretive approach will be used. The author's role as a team member of the institutional repository development project constitutes a significant basis for this decision. This is further enriched by her role as team leader for the population and marketing of the IR within the organization. Although some of the research questions can be explained through the adoption of a positivistic approach, the positivistic research methodology is not the preferred methodology for obtaining the in-depth insight required for the softer issues, namely the behaviour and reaction of the human actors. The empirical data that can be obtained by making use of the positivistic approach will provide valuable insight into existing infrastructures, as well as into the potential impact of the envisaged repository on the infrastructure. The techniques used in the positivist approach, e.g. those emphasising the observable, will help to obtain the required insight.

Use of the interpretive method falls short of providing the insights required to achieve the understanding that will be required to motivate those involved in the project. More involvement is required if the currently evident lack of compliance regarding submission and quality is to be investigated. By making use of a critical interpretive paradigm and through consultations and discussions, it is possible to identify the reasons for the lack of compliance, standardization and poor quality. It is essential to create a non-threatening environment for successful interaction. In a culture as diverse as that of the CSIR, making use solely of the interpretive paradigm cannot provide the in-depth insight required regarding the role that a diversified culture plays in the ultimate success or failure of an information system such as an IR. Critical interpretivism not only allows – but also demands – that the researcher delves more deeply into the real world of the human actors. An understanding of their fears and concerns is required. Resistance to change is a major aspect that must be included in the analysis of the final project and implementation plans. According to Ngwenyama and Lee (1997), it is essential that the researcher be sensitive to the real-life world of those involved. This sensitivity will provide invaluable inputs and insights into the investigation of questions such as lack of compliance. To date these questions, as they relate to the specific environment and context of this research project, remain unanswered.

1.7 Methodological Approach

The hypothesis is that a well-planned and structured Institution Repository using OSS will enable the management, curation and retrieval of explicit knowledge within a SET environment. It is expected that the information contained within this dissertation will contribute to the IS field by highlighting the roles that project management, systems development and management, change management and organizational cultures play in determining the outcome of an IS project.

Because of the nature of Institutional Repositories and the impact of open access publications, extensive use has been made of web-based searches. The scholarly federated search engine of Google (<http://scholar.google.com>) has been used for Internet-based literature searches, as it is also an open URL resolver. Google Scholar provides an environment whereby the researcher has access to full-text items published in scholarly journals. It also provides an overview of peer-reviewed publications. As the basic principles of institutional repositories are closely related to peer-reviewed, free and accessible web-based information, this was deemed the most suitable vehicle for obtaining current and reliable information. However, where information via Google Scholar does not meet the requirements, subscription-based sources, such as ISI Web of Knowledge, and specialised databases were used. The project itself formed a crucial part of this dissertation and a critical review is provided to determine whether the correct approach and sequence of activities were used. Errors in judgement will be noted and included as these formed a crucial part of the research.

An extensive and up-to-date literature survey was required to achieve an understanding of the current state of affairs, both nationally and internationally. The use of alerting services facilitated developed awareness of changes to technology and of the emergence of new technologies, e.g. software, thereby supplementing the existing information available. Understanding of how other institutions make use of the available technology was necessary to obtain a clear and less subjective perspective. This was achieved by joining a Listserv (Smith, 2007a) of IR developers and through close interaction with the Academic Information Services of the University of Pretoria.

As issues such as copyright, curation and ownership have specific legal implications, it was necessary to keep informed regarding possible changes in national and international legislation. To achieve this goal, the legal services of the

University of Pretoria etd – van der Merwe, A (2008)

organization were consulted to provide advice and guidance where required. The combination of literature and legal advice should therefore provide the background and insight required. The work done by SHERPA in terms of obtaining copyright clearance proved to be of immeasurable value.

As a literature survey can only supply a foundation, interviews and personal discussions with stakeholders were used to obtain a more comprehensive insight into expectations and needs. Continuous interaction was required to ensure that all the parties reached consensus. The preliminary selection of role players included representatives of the ICT team, Information Management Services, the organizational executive management team and the contributors and operators. Formal and informal discussions with selected representatives from each sector were used to determine the success of the project.

An analysis of the existing proprietary database, of identified shortcomings and of already expressed improvement requirements were used as basis for the development of the repository. The structure of the existing database was compared with IR structures currently in use. The best features of these were evaluated in terms of functionality and applicability. However, the final structure of the repository was determined by the existing limitations of the software. The assistance of the ICT development team was crucial during the development and implementation of the IR project. A relationship based on trust and respect was developed in order to develop a product that met the basic requirements of all the interested parties.

Interaction and consultation with the executive management team was required to ensure its support for the information services management team. Extensive use was made of emails to ensure that observations, interpretations and conclusions were correct. Personal contact was limited because of time constraints and conflicts in schedules. In lieu of a formal questionnaire, systems-generated statistics were used to determine the success of the project. Further interaction with the most prolific report-writing researchers contributed to identification of their frustrations and concerns. These were addressed and allayed where possible.

Discussions with the individual operators were approached on a more personal and mostly informal manner in the form of *ad hoc* problem solving, brainstorming and training sessions. Through formal meetings (CSIRIS, 2007) and informal discussions, perceived and real problems with procedures and policies were

identified and resolved. Informal group discussions were especially helpful for obtaining clarity and resolving misunderstandings.

The research methodology was based on insight and information gained by means of a postgraduate course presented by the University of Pretoria (Roode & Byrne, 2006).

1.8 Design and Layout

The theoretical considerations and research methodology were discussed in Sections 1.5 – 1.7. In order to provide a background and to move towards a shared understanding of IRs, an in-depth literature survey was conducted. The results of this are given in Chapter 2. This survey covers all the essential aspects associated with the planning, development and functioning of an information system such as an IR. Chapter 3 covers the systems development of the repository as well as the policies and workflow processes developed. Chapter 3, a case study, also provides feedback of the usage during the first four months following its implementation in August 2007 and identifies some of the development challenges. In Chapter 4, the lessons learnt are discussed, as well as the corrective action taken during the development and implementation process. Chapter 4 concludes with a brief discussion of the recommendations made, the anticipated contribution to knowledge and of further research required. Chapter 5 provides a summary and conclusion of the work done.

1.9 Conclusion

The main points of this dissertation are the development and implementation of an OSS information system (a repository) to manage, curate and distribute the explicit knowledge of an organization. Although the study is closely related to Information Science, the emphasis is on the development of the system itself rather than on the indexing and information management skills required. However, in order to place the features and functionality of the IS in context, reference is made to those issues normally associated with Information Science. It is important to emphasize the close collaboration and, in the context of this dissertation, the interdependency between the research areas of Information Science and Information Technology.

University of Pretoria etd – van der Merwe, A (2008)

Institutional repositories are continually presenting new challenges for the developers of IS products. An IR should comply not only with the demands of OSS but also with those of open/free access. With OSS still in its infancy, its development provides challenges to the developers in terms of finding the best possible products to meet their clients' demands. It is also a challenge to identify a product that will be compatible with the existing infrastructure and with enterprise-based systems. Existing proprietary systems that meet these demands are available at a cost but, as these are neither OSS- nor open-source compliant, use of these systems is prohibited and therefore a workable alternative must be identified and implemented.

Implementation of an IR in an organization calls for effective change management, targeting the developers as well as the end-users. As any IS system cannot function in isolation, the impact and demand of peripheral study areas influence the final product. If the purpose and functionality of the repository is not clear, it will be impossible to select a suitable platform. In the context of this dissertation it is impractical to separate these two study areas. The literature overview provided in Chapter 2 aims to provide the insight required and to provide motivation for linking the two subject areas.

2 LITERATURE REVIEW

Shared understanding of all the issues surrounding an IR, both in terms of an IS as well as in terms of the role of Information Science is essential. The focus of this chapter is the discussion and analysis of the information that is available in the literature. Section 2.1.1 consists of an introduction to the development and relevance of institutional repositories from a holistic viewpoint. Sections 2.1.2 to 2.1.5 focus more on the IS issues, namely on the technology itself, e.g. hardware and software. Attention is also given to the problems surrounding the preservation of digital items. Section 2.1.6 focuses on policy issues including legal implications and combines the IS side with the Information Science influence. In Sections 2.1.7 to 2.1.10, operational issues of the repository itself, e.g. preservation, compliance, quality control and auditing, and operational issues are discussed. Through-out this study, attention will be given to the human factor. It is not feasible to separate soft (behavioural) and hard (technology) issues, as doing so would create a skewed perspective. In Section 2.2 the gap between the problem statement and available literature is discussed.

2.1 Relevant Literature

Literature will firstly be discussed within the framework of the relevancy of institutional repositories. Secondly, the gap existing between the problem statement and the availability of relevant information will be identified and discussed.

2.1.1 *Introduction to the relevancy of institutional repositories*

The issues discussed in this section include an explanation and discussion of the concept of institutional repositories. The development and planning of an IR, using a worksheet, will also be discussed. Other discussions include technological issues, policies, cost models, preservation, compliance and content recruitment, as well as operational issues.

2.1.1.1 Defining the concept 'Institutional Repository'

It is necessary to understand exactly what the term 'institutional repository' means to stakeholders. The term repository as it refers to a storage unit is in itself well-known (Fowler, Fowler & Thompson, 1995). The question is often raised whether or not the term 'institutional repository' is just another name for an old and existing service, be

University of Pretoria etd – van der Merwe, A (2008)

it a library, warehouse, database, or archive. The intention is therefore to provide enough information to distinguish between an institutional repository and already existing repositories as mentioned above. The question that needs to be answered is whether it is as simple as Rankin (2005) states, namely, that an IR is ‘... a set of services for storing and making available digital research materials created by an institution’. Rankin’s definition is enforced by that of Lynch (2003) who defines IRs as services that are provided to the members of a community for the ‘... management and dissemination of digital material created by the institution and its community members.’ It is important to note that these two authors are placing an emphasis on the concept of service rather than on a physical storage area or unit. The service that is provided refers to the information system and the digital submission and preservation of information. The importance of this distinction will become clear throughout this study. In February 2008 new terms were introduced, namely the concept of a ‘digital repository’ (McHugh et al., 2007) and ‘trusted digital repositories’ (Harmsen, 2008). The basic concept is the same. However, much greater emphasis is placed 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’. Crow’s definition also focuses on a service rather than a physical storage area. He justifies his definition by pointing out that institutions, such as universities, need to take back and retain the ownership and control of scholarly communications and to reduce the monopoly currently exercised and enjoyed by publishers. However, although the emphasis is on service, sufficient and effective digital storage in terms of hardware is assumed and will not receive any in-depth discussion in this study, as the emphasis is on software and IS policies.

IRs can also serve as tangible indicators of research quality and applicability by means of usage statistics. IRs therefore have an important role to play in terms of increasing public awareness of the value of the research done by individual institutions and individuals. Crow’s (2002) reference to a multi-university community also pulls in the concept of a global community, highlighting the fact that the digital age has an even bigger impact on the information overload than did the industrial revolution of the 18th and 19th centuries or the information age of the 20th century. The term ‘university’ can, however, be regarded as a generic representation of any organization executing research, as the principles and values surrounding IRs are the same, irrespective of the core business of the organization.

2.1.1.2 The emergence of Institutional Repositories

The emergence of Institutional Repositories followed closely on the heels of the Open Source Software and Open Access initiatives. In Chapter 1, mention was made of the BOAI and the Berlin Declaration. The BOAI (2007) emphasizes the fact that the Internet had a fundamental and dramatic influence on the availability and distribution of scholarly information. This statement is underscored by the Berlin Declaration (Max Planck Society, 2003), which states that the Internet should be seen as a tool for improving the 'global scientific knowledge base'. The Declaration goes on to emphasize that all stakeholders should specify the measures that should be taken to make effective use of the Internet. Two conditions that must be satisfied in order to make open access contributions a reality are listed in the Declaration. These are:

- 'The author(s) and right holder(s) of such contributions grant(s) to all users a free, irrevocable, worldwide, right of access to, and a licence to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship ... as well as the right to make small numbers of printed copies for their personal use.
- A complete version of the work and all supplemental materials, including a copy of the permission stated above, in an appropriate standard electronic format is deposited (and thus published) in at least one online repository using suitable technical standards ... that is supported and maintained by ... well established organization that seeks to enable open access, unrestricted distribution, interoperability, and long-term archiving' (Max Planck Society, 2003).

The purpose of an IR is therefore to facilitate and meet the demands as expressed by BOAI and the Berlin Declaration. Since the implementation of IRs, open access and how this term should be interpreted became an important discussion point, as access forms the basis of any IR.

It is again important to point out that free access does not necessarily imply that access to the full text item is free-of-charge. In general, free and full access to the full text item is assumed and is within the principle of open access. Statistics presented indicate that about 74% of repositories follow the open-access approach, the remainder having some form of access control. There is a bigger trend in developing countries towards access control but about 66% of repositories in

University of Pretoria etd – van der Merwe, A (2008)

developing countries still opt for free and open access (Primary Research Group, 2007). However, this issue is still being debated heatedly by publishers, authors and knowledge organizations, as there are hidden costs that must be provided for, especially in terms of sustainability. The final decision is therefore left to the discretion of individual institutions and authors to decide whether charges will be levied for the downloading of the full text items of which they are the copyright holders.

The situation is further complicated by contractual obligations when research is done for a third party under contract. Where intellectual property rights differ from copyright ownership, separate negotiations are called for to obtain the necessary permission to publish the full text item in a repository. For the purpose of this document it is, however, assumed that free access to information includes free (gratis) access to the full text item, with no membership fees or other hidden costs, other than the users' own direct Internet and online costs. The assumption is therefore that the hosting organization will carry all the development, maintenance and sustainability costs and that access will be granted as stipulated within the Berlin Declaration (Max Planck Society, 2003).

Another reason for the emergence of IRs is the exorbitantly high costs of scholarly journals. In the current international 'library' financial environment, budgets for published journals are generally insufficient. This can be attributed to library budgets becoming increasingly limited while the costs of scientific and scholarly journals are increasing by an estimated 15% annually (Halland, 2007). In recent times, this has caused a drastic cutback in journal subscriptions, resulting in a ripple effect where end researchers are unable to keep up-to-date with what is happening in their research environments. Anuradha (2005:169) explains this by stating that IRs '... were born out of problems with the current scholarly communication model developed by commercial publishers and vendors. The establishment of IR in the developing countries ensures that their national research becomes part of the mainstream and contributes on an equal footing to the global knowledge pool'. IRs provide a workable solution and alternative to the ever-increasing costs of scholarly publications. By making information readily available, insight into the quality and value of research is provided. A further positive outflow is that IRs enable interested parties to establish contact with each other, thereby playing the role of an invisible college or community. IRs can also serve as a valuable tool for safeguarding the

explicit organizational memory in ways that a traditional paper-based archive cannot easily achieve.

In conclusion, it is essential to note that the access referred to implies Internet access, open to any interested party, and that is it generally free of charge to the end user. It also implies that OSS products will be used as these help to keep the costs lower, making the service more affordable as licence fees are excluded. The development of the information system in terms of system design forms a crucial element of the repository. Furthermore, although free and complete sharing is implied, national and international copyright laws must be observed. The final decision on whether or not to implement an IR will be influenced by how an organization can benefit from such a service.

2.1.1.3 Benefits and value of an IR

Allard et al. (2005:170) explain the value of IRs as services that ‘... provide members of the university community with the ability to add, or self-archive, items they have authored into the repository, thereby facilitating instant access to their work.’ Anuradha (2005) emphasises that a successful IR is dependent on the collaboration and cooperation between the generators of the materials, e.g. researchers and knowledge workers, and the expertise obtained from librarians, archivist, record managers, administrators, policy makers and IT staff. These statements are applicable to all generators of explicit research output, irrelevant of their core business.

The effective and clear communication of the strategic benefits of IRs is essential to ensure success and support. An adequate awareness of potential problem areas or of existing author reluctance to participate prior to the implementation of an IR is required and in this instance knowledge-sharing becomes an essential tool. One of the most important benefits of an IR is the affordable preservation and dissemination of scholarly communications. Crow (2002:6-7) highlights the fact that IRs increase access to research information, thereby also increasing competition amongst publishers, whilst simultaneously decreasing their monopolies. The strong reaction towards ever-increasing journal prices and the subsequent cancellation of subscriptions led to a move from ‘just-in-case’ to ‘just-in-time’ provision of information. It is especially here that emerging IRs address the pressing need by creating awareness, self-archiving and ‘perpetual’ access to scientific knowledge. A Unique [Uniform] Resource Identifier (URI) is used to identify and locate an item.

University of Pretoria etd – van der Merwe, A (2008)

Broeder et al. (2006:8) compare URIs with ISBN numbers, as each URI uniquely identifies an item, irrespective of where it is located. In addition, URIs are not subject to the same instabilities associated with URLs in terms of system changes.

The development and maintenance of an IR can be very expensive and will require a long-term commitment for financial support (Gibbons, 2004). An international survey done in 2006 found that the mean start-up cost of an IR in a developed country is R 12 347 990.65 and for developing countries R 567 531.12. The global annual mean operational cost is R 767 971.31 (based on Bailey et al., 2006). It is essential that the repository be defined clearly and according to the expressed requirements and purposes of the institution rather than according to pre-existing perceptions and vague generalities. The savings that use of an OSS brings is counterbalanced by the development costs but will prove substantial in the long-term.

2.1.1.4 Generic features of institutional repositories

The features of an IR are influenced by the organizational definition. However, as evident in the available literature (Anuradha, 2005; Barton & Waters, 2004; Crow, 2002; Devakos, 2006) the main challenge of an IR is to have the right technology and the required policies in place to ensure the effective long-term access and distribution of information in a digital format. Several authors (Anuradha, 2005; Barton & Waters, 2004; Gibbons, 2004) place this in context by discussing the international usage of IRs. The authors point out that, although institutions have unique cultures and assets, all tertiary institutions ultimately use IRs for the same purpose, namely for promoting scholarly communication by storing study material and courseware or by the electronic publishing of selected items. This entails the management of research and scholarly documents, including the long-term preservation of digital format. All this helps to increase the prestige of the institution by highlighting the scientific research potential of the institution.

In addition, IRs are interoperable and encourage free, open and easy access to explicit research outputs. The library plays an important role in housing and managing these digital artefacts, as it is an extension of existing services. It is important to remember that IRs are institutionally defined, community-driven, community-focussed, and are supported by the owning institution.

University of Pretoria etd – van der Merwe, A (2008)

Although current literature mainly concentrates on tertiary institutions, the basic criteria are also applicable to any organization that generates research outputs requiring long-term value that requires digital preservation and curation.

In summary therefore, the core features of an IR are to promote scholarly communication by providing access to stored digital material while ensuring perpetual preservation of material within a community-focussed framework. However, development and implementation of an information system such as a repository require a detailed analysis of the purpose, intent and goal of the repository.

2.1.1.5 Stakeholders

Based on the work of Jones et al. (2006), the following represent the current interest of stakeholders in an IR within the SET environment:

- Author: The value of future research uses, including reuse of research already done in new publications; tangible recognition, both within the organization and from funding organizations; peer recognition on an international level; long-term preservation and accessibility of the explicit output; and acknowledgement of integrity of the work done by a researcher. The calculation of the *h-index* therefore becomes of utmost importance to the authors.
- Organization: the ability to use and reuse contributions in similar environments; recognition on both national and international levels regarding the quality of research undertaken at the institution facilitated by the dissemination of the information; acknowledgement of the integrity of the work done by the organization; and long-term preservation and accessibility of the explicit output.
- Users – the scientific community: the ability to use the work in personal or new research; re-engineering of previous research; affordable access and use; easy access from remote locations; reliable/reputable information; and ensured long-term preservation and accessibility of the explicit output. The scientific community includes other research organizations, academic institutions, interested companies, organizations and interested individuals.

2.1.2 *Development and planning of an IR*

The emergence of IRs, their value and their stakeholders were discussed in Section 2.1.1.2. The most important concept of all of this is captured in the term 'institutionally defined and supported' or 'community focused framework' (Anuradha,

University of Pretoria etd – van der Merwe, A (2008)

2005; Billings, 2005; CSIR, 2007a). These terms are explained in a variety of ways within the literature. The assumption is that the needs and expectations of institutions differ and that different approaches are thus required. Differences in approach will affect the final structure, the contents that will be included, compliance management approach and the willingness to support the IR.

Information in existing publishing models is scattered throughout thousands of scholarly, peer-reviewed journals, thus making it difficult to obtain a holistic view of an individual institution's research. With IR, the otherwise scattered information can be grouped within a single information system that provides a holistic overview of the research and research quality in a simplified manner (Crow, 2002). The provision of a single entry access point for multiple repositories within a single organization also provides the holistic overview that is required. Together with technological developments, e.g. digital publishing and omnipresent networking, and the emergence of open-source standards, a significant drop in on-line storage costs has significantly stimulated the development of this much-needed service (Lynch, 2003). Lynch points out that, by centralising the stewardship of the scholarly communications, researchers can focus on the aspects in which they excel, namely research, and move away from administrative red tape (Lynch, 2003). Freeing researchers and scientists in this way contributes towards the sustainability of a repository, as they then have a stake in the product.

Crow (2002) highlights the developmental advantages of IRs at the hand of a content layer and a service layer. The advantage of separating the content and service layer is that the separation allows the institution to take control over the value-added service without infringing on the rights of individual authors. In terms of content, IRs can provide seamless searches, facilitating interdisciplinary research and discovery, a personalised content management system and the empowerment of researchers by providing a self-archiving system. Billings (2005) expands on the list of value-added services by including the link between the contents of an IR and scholarly publications of all types, including patents. All these sources could be centralised in a single location or else accessed from a single access point. The use of filters would facilitate alerting services in terms of new research in a user-specified research area (Crow, 2002) and would therefore help to prevent an information overload.

The service layer, on the other hand, covers registration, certification, citation linking and awareness and is more likely to be managed and supported by the ICT group rather than by the IR team or the relevant authors. The maintenance of file servers and network connectivity, as well as upgrading form a crucial element in the long-term availability of the IR.

For the purpose of this study, it will be assumed that the support is in place, that the IR is sustainable and that the system as a whole complies with the requirements of the stakeholders.

2.1.3 Development worksheet

With their publication, Barton and Waters (2004) provide a basic guideline document that is very valuable to any developer and planner of an IR, as it forms a checklist of all the essential activities. Their work forms the basis of this section, as they provide valuable guidance through the provision of worksheets. The authors cover issues such as development of the IR, planning of the service, selection of the correct software platform, legal and regulatory issues and cost models. The work of Lynch and Lippincott (2005) and of Rankin (2005) also serves to place the development of institutional repositories in context, although not to the same extent as that of Barton and Waters (2004). The development of an IR follows a predictable course, e.g. obtaining background information, the development of a service definition and service plan, the assembling of a development team, selection and installation of the most suitable software platform, and finally the marketing and launching of the service (Barton & Waters, 2004).

Some pressing challenges face the development and implementation of an IR. Included in these is the rate at which the IR is adopted by knowledge workers and at which compliance improves within and according to internal policies. The provision by the institution of a sustainable infrastructure is crucial. Efficient management and recording of intellectual rights and copyright issues are essential to protect the integrity of the institution. The availability of institutional support, cost management; digital preservation and identification of all the stakeholders are other challenges that must be resolved to ensure the success of the repository (Barton & Waters, 2004).

University of Pretoria etd – van der Merwe, A (2008)

The workbook prepared by Barton and Waters also provides valuable guidelines for defining the purpose and services of the IR, in terms of both contributors and end users. Some of the issues they regard as essential represent a clear statement of the mission of the IR. Others include the identification of key stakeholders, affordable services and top priorities, from short-term to long term (Barton & Waters, 2004).

Anuradha (2005) provides valuable insight to this process by presenting her analysis of the implementation of the institutional repository, PRABHAVIS, at the Indian Institute of Science (IISc). The works of some other authors, e.g. Björk (2004), focus on the barriers or resistance to change with which an IR are confronted and provide some insight on how to address these issues. Rankin (2005) provides an extensive framework of topics that should be considered during the development, planning and implementation of an IR. These include issues such as the contents of the repository, the set-up and maintenance costs of the repository, including the availability of off-the-shelf software vs. the development of an OSS product.

Other issues listed by Rankin that influence the individual institutional definition include organizational, administrative and cultural issues. He also includes content policies, accession retention and preservation of the digital contents issues. The involvement and participation of the researchers, intellectual property rights and access rights should also receive attention. Lastly, it is important that technical, technological and infrastructure issues be considered (Rankin, 2005).

The preservation issue should be addressed as part of the development plans of the IR (see Section 2.1.8). Although persistent URIs do enable long-term access, media-dependent data will require migration from time to time to ensure long-term preservation (Billings, 2005). However, with the implementation of an IR, the responsibility and costs for the migration process are transferred from individual authors to the organization, i.e. the IR administrators, thereby facilitating the long-term access and availability of data. Because IRs have the potential to become tangible indicators of research relevance and activities, long-term access to the data is essential (Crow, 2002). This emphasises the need for high quality and durable migrated information in order to ensure long-term retrieval. In this context, quality and durability refer to the digital format of the item and not to its content. However, even more challenges face the IR team.

2.1.4 *Challenges facing implementation of an IR*

Although Björk (2004) tends to view the situation from an open-access viewpoint, his principles are applicable to all IRs. The competition between an IR and commercial publishers is a major obstacle that cannot easily be overcome. Other typical barriers are the legal framework within which the organization functions, existing organizational IT infrastructures, business models, recognition and awards, marketing, critical mass and the standard of indexing services. It is impossible to provide a ready-made solution as the character of the organization will determine the route to be taken. Awareness of potential pitfalls is therefore essential during the planning and development process.

Van Westrienen and Lynch (2005) report on the status of academic IRs as it was in 2005 and Sanger (2006) looks into the future with his concept of the perfect information resource. Although Van Westrienen and Lynch did not execute a truly global survey, they do provide a glimpse of the situation as it was in the early stages. They list 35 repositories in 12 countries. By 2007, this number had risen to 56 repositories in 11 countries surveyed (Primary Research Group, 2007). Sanger's perfect information resource is fully aligned with the concept of an IR. His description of the resource can be summarised as follows: An Internet-mediated project featuring maximum involvement by both experts and the public, working together and free of commercial influence providing a maximally free product that is arranged into taxonomically sorted portals (Sanger, 2006:4). If Van Westrienen and Lynch's information is compared with Sanger's perfect resource, it is clear that IRs still have a long way to go to address the needs existing within user communities.

As information relating specifically to the scientific community is scarce, the insight provided by Anuradha is invaluable, especially her tabular analysis of the identified metadata fields used in the PRABHAVIS repository (Anuradha, 2005). In addition, the author points out that the data were harvested from a variety of sources, resulting in a variety of formats, standards and in duplicate records. This is a common occurrence when using harvested data. The value of making use of harvested data vs. the need to validate and monitor the content is thus debatable. During the development of PRABHAVIS, priority was given to the collection, archiving and dissemination of information of the data. However, the developers at IISc went further to determine that the IR should provide a single access point and should act as a self-evaluation tool (Anuradha, 2005). In order to achieve their goals,

University of Pretoria etd – van der Merwe, A (2008)

the developers had to implement algorithms to sift the records in order to standardise the metadata and to remove duplicate records. Although data mining falls outside the scope of this study, it deserves future analysis and monitoring. Anuradha (2005) provides a brief background on how these algorithms were developed and implemented. However, the background provided by Anuradha will be useful to any institution planning to use data mining (harvested data) and for a theoretical comparison of existing database vs. that which is required for an institutional repository.

During the development of the IR, some critical core functions should be planned and provided for. A mistake here could be very costly and could lead to the ultimate rejection of the IR. The following are loosely based on the work of Billings (2005):

- Submission and editing of digital material by the author or other authorised individuals – how will internal control authorise individual rights to add or edit items?
- Enhancement of the metadata, e.g. authors, keywords, abstract and physical description by implementing standards and guidelines – the monitoring of quality to ensure user satisfaction is essential and must be managed. Is it possible to implement internal controls and to what extent can this process be automated?
- Management of access rights including the right to add, edit, delete and approve the records, will change over time and an authorised or empowered party should be appointed and detailed activity logs kept. Detailed logs are required to monitor this activity and internal controls are required to allow authorised changes only.
- Registration of the IR with search engines and service providers such as Google, OAIster, and DOAR (Directory of Open Access Repositories) - an increase in visibility will contribute to the value and sustainability of the repository. The IR manager must keep up-to-date regarding service providers and automated web crawling will help to identify the most suitable search engines and service providers.
- Implementation of mechanisms and action plans for the long-term preservation of the records and upgrading of the digital formats. The assistance and support of the organizational ICT group will form an essential and integral part of ensuring long-term preservation, as it is a complex, time-consuming and labour-intensive process (See Section 2.1.8). The question remains of how the IS can help to identify those items that are in danger of becoming obsolete and whether it is at all feasible to try and implement this. What about quality control and

University of Pretoria etd – van der Merwe, A (2008)

preventative measures that must be in place to prevent malicious or accidental corruption or loss of data?

Authors are generally in agreement regarding the basic scholarly content that should be included in IRs. In general, scholarly content includes peer-reviewed journal articles and conference papers, data sets, theses and dissertations, book chapters and grey literature such as research reports (Anuradha, 2005; Barton & Waters, 2004; Billings, 2005; Gibbons, 2004; Jenkins, Breakstone & Hixson, 2005; Johnson, 2002; Lynch & Lippincott, 2005). It is important to note that the institutional needs will define the scholarly content in more detail than is provided here, especially in terms of contractual obligations. For instance, universities will typically include classroom teaching materials and study reserves. On the other hand, a SET organization might include selected contract reports and publications funded by government departments. Lynch and Lippincott (2005) provide an extensive list of content types typically included in IRs. The distribution in terms of content types worldwide is illustrated in Figure 1. It should, however, be emphasised that content types are subject to interpretation, e.g. conference and workshop papers may be subjected to additional criteria, namely peer-reviewed items vs. non-peer reviewed items. In addition, it is not clear exactly which content types include articles published in peer-reviewed journals although this is normally associated with the category 'Research papers'. It is clear that there are still some room for improvement and refinement in the definition of the information provided. The technology selected should make provision for all possible content types and formats. One of the challenges is to identify and define the content/publication types that will be included in the repository.

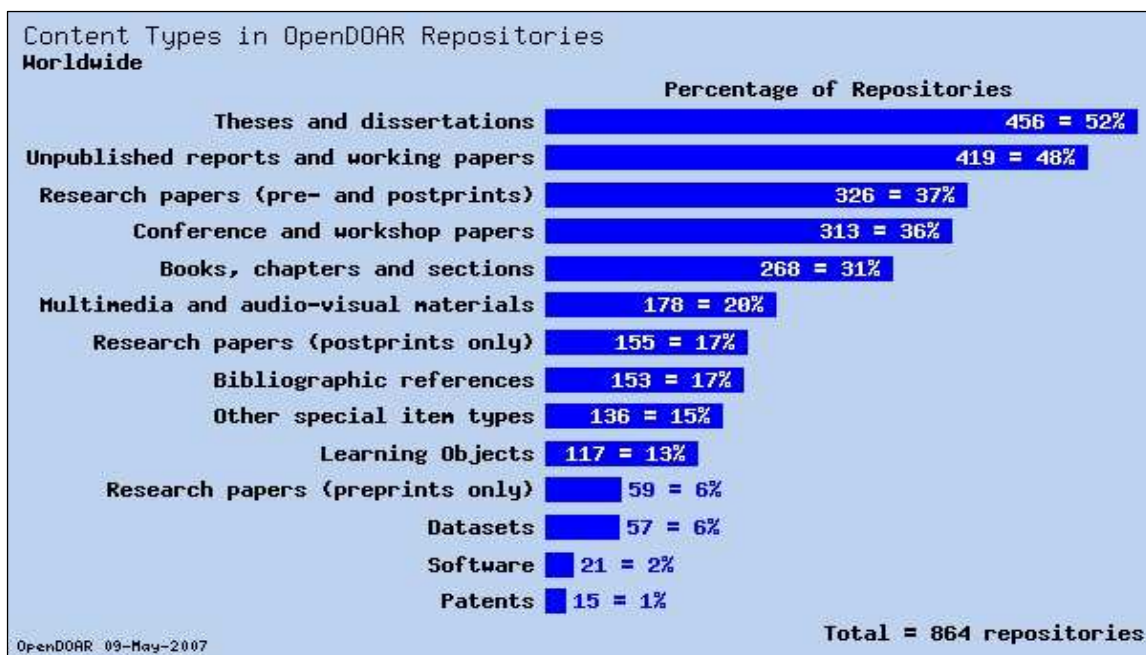


Figure 1: Content types as reflected in OpenDOAR repositories (OpenDOAR, 2007a)

2.1.5 Technological issues

Boulanger (2005) provides an interesting comparison of proprietary and Free Open-Source Software (FOSS). FOSS was developed by the informal collaboration of volunteer programmers and the proprietary software was developed by a formal and dedicated design team consisting of designers, programmers, project managers, and quality assurance engineers. In stark contrast with proprietary systems, FOSS systems themselves, excluding support, are provided free of charge (Boulanger, 2005). Support for FOSS is provided either free-of-charge or for a nominal fee or by subscriptions from the programmers themselves. In general, the developers of the software provide support for proprietary software and it takes place in a much more structured manner and within the context of ownership.

One issue that is hotly debated between proprietary software supporters and FOSS supporters is that of security. One of the foundations of the debate is the general availability, or lack thereof, of the source code of the software. Boulanger (2005) points out that the unavailability of the source code does not guarantee security, as the source code is not required to locate vulnerabilities. Although the availability of the source code also does not guarantee security, it makes it easier for the FOSS community to analyse the software, identify the weak spots or defects, identify

malicious behaviour and, subsequently, to take corrective action. The corrective action is shared with the community, which can then take action, rather than having to wait for the vendor to supply the required patches (Boulanger, 2005). This debate is still far from being resolved and laid to rest. Boulanger points out that ‘...FOSS systems can meet, or even exceed, the quality of their proprietary counterparts ..’ and that ‘FOSS-developed systems offer viable alternatives to proprietary systems in terms of software quality and reliability’ (Boulanger, 2005:245). Proponents of proprietary and FOSS products continue to argue the benefits of their individual approaches and it remains the prerogative of the user to select the most suitable product.

2.1.5.1 Proprietary vs. Open Source Software

As stated earlier, the advantages of open source software (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. Most research organizations, such as the CSIR (South Africa), used proprietary systems prior to the decision to move towards OSS. The challenge that now emerges is to find a suitable software product and then to convince the users that the move is feasible in both the short and long terms. As can be seen from the comparison in Table 1, OSS products have most of the essential features required; however, some compromises will have to be made when selecting the most suitable product. The comparison provided here is between the current proprietary InMagic system (InMagic, n.d.) and the OSS system, DSpace (DSpace Foundation, 2007) as the potential OSS product for the implementation of an IR. The value assigned to each feature was determined at a meeting of information specialists and indexers as indicative of the ideal situation. Fact sheets supplied by the distributors, as well as personal experimentation and experience were used for the comparisons. Where a specific functionality was available, a manual comparison was done. For example, the Validation list was compared in terms of both creating and modifying data as well as the actual end-user usage.

Table 1: Comparison of functionalities - InMagic (proprietary) and DSpace (OSS)

| Functionality | Value | InMagic | DSpace |
|---|-------|---------|--------|
| Validation lists/controlled vocabulary | E | A | L |
| Auto spell checker | E | A | A |
| Unique fields | E | A | D |
| Multiple entries fields | E | A | A |
| Easy manipulation of structure | R | A | P |
| Easy creation of forms – displays, reports etc | R | A | P |
| Export in ASCII/XML | R | A | P |
| Log sheets | E | A | D |
| Statistics | E | A | A |
| Web enabled | E | A | A |
| Easy maintenance, i.e. upgrades | E | A | P |
| Support | R | A | P |
| Strong search functionality | | | |
| Boolean searching within fields | E | A | L |
| Boolean searching between fields | E | A | D |
| Combination of the above | E | A | D |
| Truncation – end of word | E | A | D |
| Truncation – middle of word | O | N | D |
| Truncation – beginning of word | O | N | D |
| Guided search functionality | E | A | D |
| String searching | O | A | D |
| Complex searching – long strings | R | A | D |
| Sorting of results | O | A | D |
| Email functionality – search results | R | A | D |
| Canned searches | O | A | D |
| Output formats | | | |
| RTF | R | A | D |
| XML | R | A | D |
| Text | R | A | D |
| HTML | R | A | D |
| Spreadsheet | R | A | D |
| Browse function | R | A | A |
| Personalization of query screens by user | R | A | N |
| Variety of input screens according to publication types | R | A | A |
| Ability to create bullets or paragraphs in abstract | O | A | A |
| Calculations functionality, e.g. publication equivalency scores | O | A | N/D |
| Linking to full text – variety of formats | E | A | A |
| Searching of full text | R | N | N |
| Unlimited users – search | E | A | A |
| Input – limited to selected users | E | A | A |
| Limited access/protected fields/hidden fields | R | A | A |
| Controlled access – adding/editing/deleting | E | A | A |
| Must not be limited in the number of records that can be added | E | A | A |
| Must have a date field where the last update is indicated | E | A | A |

Value: E=Essential functionality; R=required but can be developed; O=optional (nice to have); Software: A=available in 'vanilla format'; D= not available, requires development and will result in additional cost and time; L=available but not on the level required, will require development resulting in additional cost and time; P=specialized programming skills required; N=not available.

University of Pretoria etd – van der Merwe, A (2008)

As drive towards OSS continues, perceived problems are resolved and usage increases. Users will accept that OSS has an unqualified right to existence. However, it might be necessary to sacrifice some of the functionality to stay within the national drive to move toward OSS products.

In general, an organization tends to ask for unnecessary enhancements and add-ons because they look good on paper. By utilizing the techniques and methodologies of the critical interpretive paradigm as well as empirical data, it will be possible to understand the organization's real needs and requirements. This will allow for well thought-through recommendations for, for example, suitable types of management reports and for the correct formats of these reports. It is important to note that the decision regarding the most suitable software will also be influenced by policies, cost models of the repository and preservation issues.

2.1.5.2 IR software

Technological issues are discussed in several publications. The Open Society Institute (OSI) (Open Society Institute, 2004) provides a detailed and very valuable analysis of the existing repository software, including a comparison between the two top performers, DSpace and EPrints. This is illustrated in Figure 2.

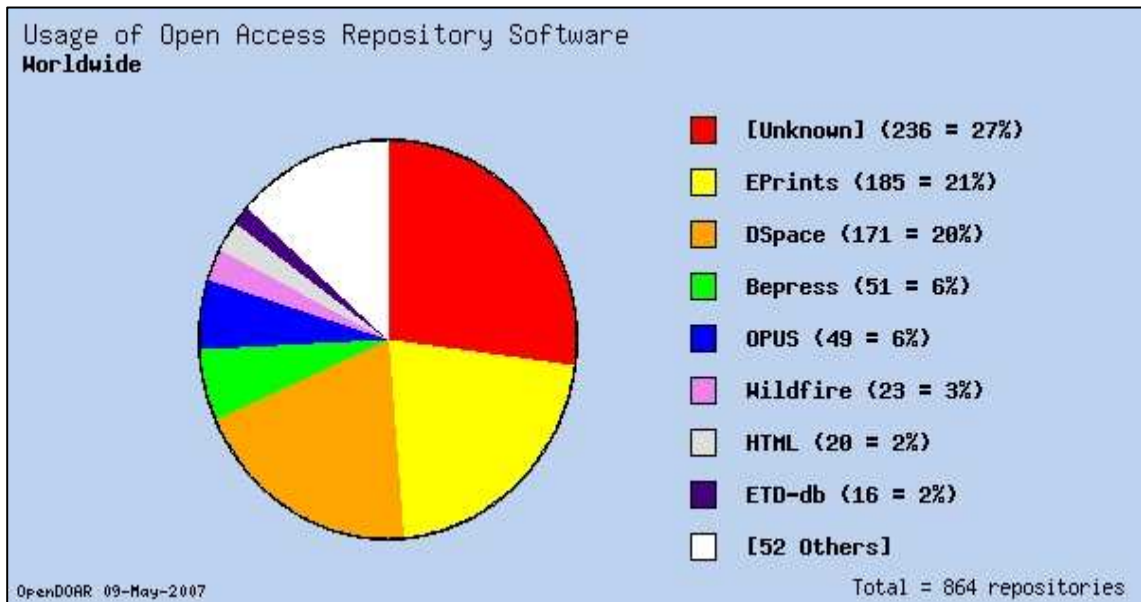


Figure 2: Graphical representation of Repository Software (OpenDOAR, 2007b)

University of Pretoria etd – van der Merwe, A (2008)

The comparison provided by OSI covers, *inter alia*, the following:

- **Technical specifications**, e.g. hardware, software, required skills on the side of the developers and support staff and the browsers that are supported;
- **Repository and system administration**, e.g. issues such as installation and user registration, authentication and password administration and submission support;
- **Content management**, i.e. issues such as acceptable formats, importing and exporting of data, metadata standards, user interfaces and search capabilities;
- **Archiving**, covering issues such as persistent document identifiers (URIs) and data preservation support (amongst others, the migration of data between different versions of the software); and,
- **System maintenance**, e.g. documentation and formal support, backup and general service level agreement (SLA) management (Open Society Institute, 2004).

DSpace (DSpace Foundation, 2007) was developed by MIT with the expressed purpose of creating a suitable platform ‘... to capture the intellectual output of multidisciplinary research organization.’ Version 1.2 of DSpace was released early 2004. DSpace has a strong user community that enriches the development of the platform. This approach to providing a versatile and adaptable product is visible in the customization of the workflows and any other issues that are dictated by individual organizational policies. According to the OSI, this makes DSpace ideally suitable for larger institutional organizations (Open Society Institute, 2004). DSpace software is freely downloadable from <http://www.dspace.org>. An international survey found that 83% of the repositories in developing countries use DSpace, by comparison with 37% of those in the rest of the world (Primary Research Group, 2007).

EPrints (EPrints, 2007) currently shows a slightly wider user base than DSpace. EPrints was developed at the University of Southampton and was released in late 2000, giving it a head start of more than three years over its greatest competitor, DSpace. The software can be installed without extensive technical expertise, allowing the IR to be up and running in a short period. Because of the growth in usage, baseline capabilities, such as an integrated advanced search option and extended metadata has shown that the product can be customised according to users’ specifications and requirements (Open Society Institute, 2004). EPrints software is freely downloadable from <http://www.eprints.org>.

Jihyun provides a comparison of finding documents in EPrints and DSpace from a user's perspective. The author describes the approach as a '... a heuristic evaluation and usability testing [measuring the] time for completing tasks, the number of errors and users' satisfaction...' (Jihyun, 2005). Both systems were evaluated using Nielsen's usability heuristics focussing on intuitive use of the two systems. The elements in Nielsen's heuristic model include aspects such as dialog and language, consistency, easy navigation, quality of the error messages and online help and documentation (as quoted in Jihyun, 2005). By using Nielsen's usability heuristics, Jihyun proposed 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 (Jihyun, 2005).

Jihyun concludes by pointing out that the DSpace interface is preferred to EPrints. However, he also pointed out that both interfaces proved difficult in some areas. The improvements required are in terms of search options, examples of queries to use as guidelines, the display of search results and links to full-text documents that are more visible within the EPrints environment (Jihyun, 2005). Ultimately, the culture and needs of the specific organization will determine which product will be most suitable to it.

Personal experience with the default search functionality of DSpace was disappointing. Shortcomings include: a) lack of in-depth Boolean search capabilities; b) inability to search for a word-in-title; c) ineffective author searching; and d) ineffective keyword searching. In most cases, the browse option proved to be more effective than either the basic or advanced search functions. Fortunately, search engines such as Google can be used effectively.

Nixon (2003) provides the reader with lessons learnt during the implementation of the DAEDALUS project, in which both EPrints and DSpace were used, and he provides a useful comparison between the two systems. According to Nixon, the installation of DSpace was more complicated than that of EPrints. However, he acknowledges that prior experience with the EPrints platform could have been a contributing factor. The University of Glasgow could attribute some of the problems experienced with the installation of DSpace to the Solaris and Tomcat versions in use at that time. Both systems can be customised significantly, including the

inclusion of logos and add-on modules and both were developed with self-archiving in mind. However, both systems can also be set up in such a way that a workflow for the moderation of submissions can be implemented. Both systems have a personalised space in which users can monitor the work submitted and most administration is managed via a Web interface. One of the differences between EPrints and DSpace is the level on which access is granted. EPrints provides access on an item level. DSpace, on the other hand, not only allows for access control on an item level, but also on community, collection, and bitstream levels (See Section 3.2.2. for an example of an implemented system). Nixon concludes by pointing out that the final selection will be influenced by local factors, e.g. existing expertise within the organization, the purpose of the repository and preservation policies (Nixon, 2003).

Beier and Velden (2004) point out a very important aspect that must be considered during the selection and implementation of the IR software, namely its compatibility with existing products. In any distributed environment, it is essential that the selected software have a proven compatibility with existing systems, especially in terms of workflow and infrastructure. This can also include reference systems such as RefWorks and Reference manager, imports and exports via XML and linking to full text items subjected to external copyright issues. This need for compatibility proved to be the deciding factor when DSpace was selected as the preferred platform for CSIR Research Space. The following section provides more information on the DSpace software and Section 3 expands on the selection of this product by CSIR.

2.1.5.3 DSpace

One of the features of DSpace is an embedded workflow. This, however, should not be confused with the workflow surrounding the document management process within which the IR falls. The document management process addresses the organization's needs to identify items for inclusion, selection of the applicable collection/s and approval of items subjected to IP authorization (See Section 3.2.2.2.).

The embedded DSpace's workflow consists of several processes. The DSpace Ingest process is illustrated in Figure 3.

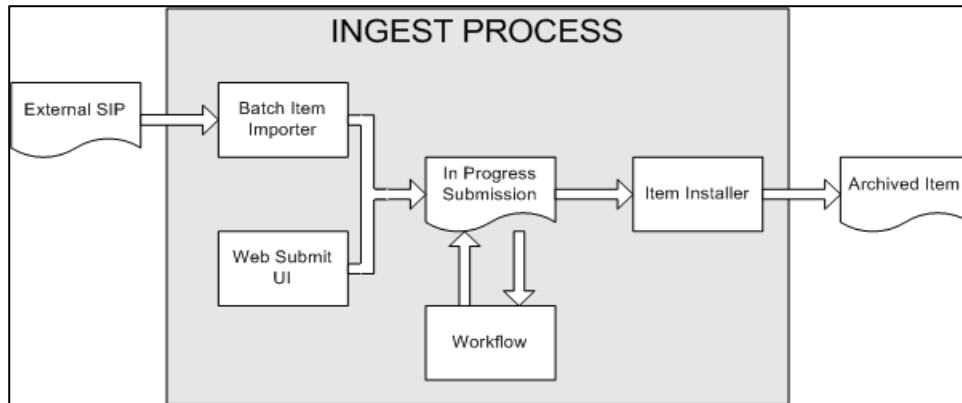


Figure 3: DSpace Ingest process (Tansley et al., 2007)

According to the developers, the batch item importer is an application. This application turns the external SIP (Submission Information Package) into a progress submission object that might in turn activate a workflow process. The workflow process generally calls for human reviewers to check the submission and to ensure that the submission is suitable for inclusion in the collection. When the 'InProgressSubmission' object is completed by the 'Batch Ingester' the next stage of the ingest process is invoked. Following this, a provenance message is added to the metadata, including filenames and checksums of the contents of the submission. Similarly, with any activity a provenance statement is added. On the successful completion of a workflow process, the object is taken up by an 'item installer', which converts the InProgressSubmission into a formal archived item. The item installer therefore assigns accession and availability dates to the metadata. An issue date is also added if this has not previously been done. Provenance messages are added, a persistent identifier (URI) assigned and the item added to the target collection. Appropriate authorization policies are added and, lastly, the item is added to the search and browse indices (Tansley et al., 2007).

The individual workflow processes associated with collection are linked to a human reviewer or gatekeeper as mentioned. It is important to note that, should the manager of the repository not have established such a link, that step in the workflow will be ignored and the item will immediately be archived and will therefore be immediately available. The generic workflow process is illustrated in Figure 4. The gatekeeper should be identified in the individual policies.

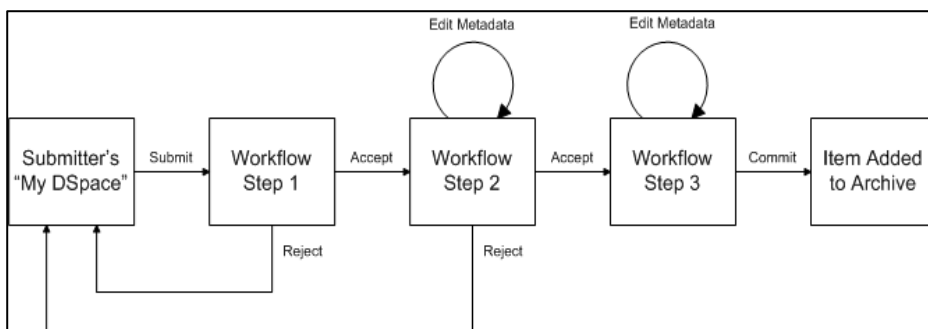


Figure 4: Submission workflow in DSpace (Tansley et al., 2007)

2.1.6 Policy issues

Those directly affected generally regard existing policies implemented by the organization as dictatorial, as they are not aware of the bigger picture. The ideal solution is to involve representatives of the knowledge workers as well as of the operators during finalization of the policies. Mutual agreement and understanding is required if the benefits are to be realised. By making use of the critical interpretive approach, the researcher will be able to anticipate potential dead ends and bottlenecks. Making timely recommendations for the implementation of other options will be possible. The identification and resolution of negative perceptions regarding policies will be required prior to implementation of the new information system. The policies should: a) include the provision of relevant system-based training programs; b) include feedback to the development team to determine whether the final product meets the goals stipulated in the original purpose statement; c) identify gatekeepers and other role-players, e.g. support personnel and submitters; and d) cover aspects such as service models, compliance issues, legal implications, cost models, service level agreements and content types. Policies will be heavily influenced by the philosophy of the organization and the ultimate purpose of the IR. One of the aspects that should be considered is the service model.

Branin points out that although the technological aspect is a crucial part of the IR service it can prove to be the ‘... least expensive and least complicated component’ (Branin, 2003:13). By linking the service model of an IR to Digital Asset Management (DAM), Branin lists six elements that should be covered in the service model. These are: a) digital asset creation and submission; b) preparation of the metadata; c) IP (intellectual property) rights management; d) preservation management; e) assistance with content access and use; and d) marketing (Branin, 2003:13). The role of Informatics is especially required for the creation and

University of Pretoria etd – van der Merwe, A (2008)

submission of digital assets, advising on preservation management and assisting with access control. Informatics can also provide valuable inputs in terms of tools and controls that can be used with metadata and IP rights management and quality.

2.1.6.1 Essential elements in policies

Barton and Waters' (2004) publication is of enormous value in the drafting of a policy. Their workbook approach covers all the important issues and serves as a valuable checklist. In addition, other authors (Anuradha, 2005; Mackie, 2004; Nixon, 2003) repeat the advice given by Barton and Waters. Although extended reference will be made to the work of Barton and Waters, cognisance was taken of the work of the other authors. The value of Barton and Waters' work is that it pulls together the essential issues that need to be considered, without being dictatorial.

According to Barton and Waters, there are three types of policies to consider. The first of these is those policies that can be resolved internally by the project team, e.g. supported formats and quality control. Secondly, there are those policies that relate to existing policies, e.g. collection types/groupings and access to the collections, e.g. paid or free. Lastly, there are those policies relating to the greater organization, e.g. user authentication and identification (in terms of access rights) and privacy policies (on an individual or organizational level) (Barton & Waters, 2004).

Barton and Waters also provide insight into how the design of policies could be approached and which topics should be considered when the policy is being drafted. These elements include a policy advisory group that can advise on issues such as submission and distribution, as well as privacy and licensing issues. According to Barton and Waters a typical advisory group will consist of representatives of IT (Information Technology), Public Services, Collection Management Services, Document Services, Archives, Divisional Libraries, IS (Information Systems), Support Services and Communications. Obviously, the composition of the advisory group will be influenced by the organization itself and its needs (Barton & Waters, 2004).

Aspects that should be considered in the policy cover the definition of the content/publication types, collections and guidance regarding legal issues (Barton & Waters, 2004). The developers require policies that determine the publication types, the acceptable formats, and access controls associated with the different types and formats of the publications. In addition, the developers need clear guidance

University of Pretoria etd – van der Merwe, A (2008)

regarding the management and administration expectations of the stakeholder. They need to be informed about the required structure of the system and about any embedded or external workflows that should be incorporated into the system, e.g. approval rights. The developers should also be aware of any copyright and legal issues that need to be managed. However, it is important to note that input is required from the Information Sciences side, as these individuals will have the primary responsibility for the final product.

Although the checklist provided is very similar to the planning checklist discussed in Section 2.1.3, more details regarding the actual questions are supplied. Although the following guidelines were based on the work of Barton and Waters (2004) it is enriched by applicable examples. These are:

- What types of material will be accepted, e.g. strictly digital formats to match the free accessibility associated with IRs or a hybrid system. If acceptance is limited to digital formats only, which types of file formats would be acceptable, e.g. *.doc, *.pdf, or others? It should be noted that this decision would affect the preservation issue as well as storage requirements.
- Whose work will be included in the repository and who will be allowed to submit the items? Although the norm is only to include works by authors affiliated to the organization, joint ventures can have an effect, especially on authority files, access rights and intellectual property rights.
- How will the repository be structured, e.g. on a departmental or individual basis or perhaps on a combination of the two options? Who will maintain and develop the structure in future? Has the system been developed in such a way that delegation is possible? Are there any contingency plans in place in case a department or group ceases to exist? The influence that a dynamic organization will have on the IR should not be underestimated. (See Section 3.2.1 for an example of how this can be managed.) It is essential that the system be flexible enough to allow for changes.
- Free access or managed fee? Access can be open, embargoed or restricted. The potential impact in terms of intellectual property rights should be considered. The result might be a hybrid free/managed system. Internal controls should be in place to manage issues such as the access, subscription renewals and to monitor payments.
- Agreement in terms of an acceptable downtime should be negotiated by means of a service level agreement.

University of Pretoria etd – van der Merwe, A (2008)

- Preservation: Questions that need to be answered are: a) which formats are supported and to what degree; b) how are version control and migration managed? c) who is responsible for backups and how will these be managed, stored, and retrieved? d) how will quality be monitored; e) how great is the risk of data corruption; and e) what is the cost model that will be used in terms of preservation. Preservation can be very expensive in terms of labour costs, technology and storage.
- Withdrawal of items: Is this allowed? What approval will be required and how will withdrawals be audited and managed? Who will be authorised to withdraw items? The developers will have to incorporate internal control measures to ensure that malicious access can be prevented.

Barton and Waters (2004) mention the value of a Memorandum of Understanding (MoU) that can serve as a tool to protect the rights of all the parties involved. Although they approach it strictly as an agreement between the University and the IR, it has additional value in terms of joint ventures and subsequent publications. The value and structure of a MoU will be determined by the organization and by the type of scholarly communications generated by the organization, its knowledge workers and the relationship between the organization and its partners.

2.1.7 *Cost models*

Costs models are another important aspect that will determine the sustainability of the IR. The costs issues to be considered include direct costs such as those relating to system development, resources, scale and service maturity. Hidden costs to be considered relate to strategic planning and support staff; office space and utilities, training and marketing and to preservation and disaster recovery.

Some of the key questions provided by Barton and Waters that affect the cost model are:

- Is there a need for additional administrative assistance to support new personnel?
- Is there a need for specialised space and/or equipment to support the IR?
- What is required to support the IR in terms of specialised IT skills, e.g. Java programming? Does existing staff require additional training to align its capabilities with the new requirements or should a new appointment be made?

University of Pretoria etd – van der Merwe, A (2008)

- What impact will the IR have on the existing ICT support staff, e.g. will additional staff be required, or can IR be accommodated within the existing workload?
- Which resources/services can be outsourced and, if so, how should they be managed?
- How can the expenses associated with the long-term preservation of digital format be planned and accounted for (Barton & Waters, 2004)?

2.1.8 *Preservation issues*

As just mentioned, long-term preservation costs should be included in the cost model. Goh et al. (2006:368) mention that preservation not only refers to the preservation of metadata but also to the use of quality control measures to ensure integrity, and persistent documentation identification for migration purposes. However Goh et al. and most of the other authors discussing IRs do not cover issues such the preservation of the digital format itself, the problems and costs associated with changing formats and version, technology becoming obsolete and outdated or software and hardware becoming incompatible with each other. Fortunately, other authors are focussing on these problems. According to Smith (2002) it is impossible for digital formats to survive or remain accessible by chance. He compares the changeability of the digital format with a volatile and fickle object and emphasises that society in general does not understand or appreciate the complexity of the problem. He goes on to warn that technology should not be allowed to determine preservation policies. Aspects that contribute to this complexity are the issues of costs and scale, the necessity of building appreciation for the problem and those relating to technical and research issues. As technical issues can be regarded as the most pressing, a clear set of priorities should be created. Priorities should include conceptual requirements, tools and technical infrastructures, standards, security and IP (Intellectual Property) Rights protection and prototypes and trials to test the models proposed. Compatibility issues when moving from one brand to another, e.g. Corel to Microsoft, provided a personal insight into the complexity involved. Personal experience has shown that any conversion should be approached with caution. Conversion is generally expensive in terms of real costs, person-hours and corruption of data.

This opinion is reinforced by Kuny (1997) who warns that society is moving towards a 'digital dark age' and that, unless proper precautions are taken, much of what is available today will be lost for future generations. Although IRs address one of the

University of Pretoria etd – van der Merwe, A (2008)

main reasons for the loss of knowledge, e.g. proper archiving rather than the lack thereof, other issues that still need to be addressed are the fact that technologies are becoming obsolete, that there is a proliferation of formats (each potentially requiring specific hardware and software), a lack of standards and costly licensing fees. IRs are currently addressing some of these issues. These include the need to develop a typology of data types and formats, examination of the various cost models that will ensure the sustainability of IRs and thereby the preservation of the digital formats and investigation of the preservation needs of research organizations, including universities. The role that Informatics must play here should not be underestimated. IT specialists are in a position to guide and advise IR managers so as to ensure that the proper planning and budgets are in place.

However, there is still a gap in terms of the three main methods of preservation listed by Kuny (1997) namely technology preservation, technology emulation and information migration. The need to find a solution is evidenced by the amount of information already lost because of the loss of the required technology – both hardware and software. Kuny says that, although digital collections facilitate access, they do not facilitate preservation, since being digital equates being ephemeral. Digital also places greater emphasis on immediate availability (just-in-time information) rather than on traditional long-term storage (just-in-case) (Kuny, 1997). IT specialists have an important role to play in moving digital formats from the ephemeral to the durable.

Action is being taken to identify and measure the risks and to provide guidelines for the preservation of these formats. Stanescu (2005) discusses the development of the INFORM methodology. The purpose of INFORM is to analyse the threats facing preservation in term of durability. It also defines the tools and evaluation process required. INFORM provides the reader with six classes of risks, namely:

- Digital object format: Risks introduced by the format specification itself and by dependent specifications of compression algorithms, proprietary (closed) vs. open formats, DRM (copy protection), 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

University of Pretoria etd – van der Merwe, A (2008)

and digital signatures. Examples include: unavailability of the source codes and incompatibility between versions. Alternatively, the source codes might depend on external libraries.

- **Hardware:** Risks introduced by necessary hardware components, including media type (CD, DVD, magnetic disk, tape, WORM), 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 the fact that the user community might not be effectively involved in preservation planning.
- **Digital archive:** Risks 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 a likelihood for corruption of data to occur; b) access security is weak, allowing unauthorised or accidental alteration or deletion; and c) off-site storage of hardware, media and software 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.

All the issues mentioned until now will become irrelevant if there is not a reasonable guarantee that the repository will be populated with the relevant information. In order to do so, it is necessary to look at the compliance and content recruitment issues associated with a repository.

2.1.9 *Compliance and content recruitment issues*

As mentioned earlier, it is not possible to discuss the development and implementation of an IR without touching on some of the issues normally associated with Information Science. Compliance is one of these issues and is included in order to develop a more holistic perspective of the issues surrounding a successful IR. Compliance with IR policies by obtaining the 'voluntary' cooperation of authors provides a challenge to the IR managers as evidenced in the work of Lynch (2003), Mackie (2004) and others. It is also one of the most challenging 'soft' issues to resolve. Some of the issues mentioned by Lynch (2003) may explain, to some degree, the resistance of authors to submit their work to an IR. Authors may regard IRs as a controlling tool, thereby removing ownership of intellectual property from individual departments.

Other authors provide the same basic explanation for the lack of participation (Jenkins, Breakstone & Hixson, 2005; Mackie, 2004; Mark & Shearer, 2006). Among these are concerns that existing relationships between authors and the publishers of peer-reviewed journals can be harmed – a 'do not bite the hand that feeds you' attitude. Authors are also concerned that contents in IR will not receive the same recognition/accreditation as items published in accredited journals, as peer review might be absent. More tangible are the concerns that there might be infringement of copyright laws, with far reaching results for both the author and the organization. There is also a legitimate concern on the side of the authors that they might lose their individual ownership by transferring it to an organization. Perhaps less valid is the prevalent reluctance to trust a third party to take care of the long-term viability and sustainability of digital content and formats. Associated with change management is the unwillingness to change existing work processes and patterns – existing comfort zones. This goes hand-in-hand with a reluctance to share draft versions of work – fear that others might 'steal' ideas and then benefit from them. The slow rate – and sometimes failure – of accredited associations to prioritize or support changes in scholarly publishing reinforces the reluctance to submit items to the repository. The monopoly of a handful of publishers in specific and pre-existing forums for sharing scholarly work – and which creates a comfort zone – compounds the concerns about accreditation. Lastly, without visible incentives, namely the age-old question of 'what is in it for me?', support of an IR will remain low. As there is some basis for these fears and concerns, they should be resolved in order to ensure the success of the repository.

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 knowledge workers/researchers are. The third step is to enhance the IR so that it meets the needs of potential stakeholders and can accommodate existing work practices. Lastly, it is essential that stakeholders understand, on a personal level, the long-term benefits of the IR. The four steps listed will require a personalised approach to address concerns and expectations. It remains clear however is that potential contributors should be aware of the benefits mentioned earlier. Mark and Shearer (2006) include marketing as a fundamental process necessary to create an awareness of the benefits of an IR. They also mention that the services provided can be extended to include an assisted depositing service whereby the metadata is added by the IR staff. The IR personnel then undertake any copyright negotiations. (This approach has been adopted at the CSIR.) Another way of obtaining content, mentioned by Mark and Shearer (2006), is the harvesting of information from any resource that allows harvesting and then incorporating the data into the IR. However, the delays associated with this approach and potential impacts must be evaluated before taking this route. On the other hand, this is a very effective approach to populate repositories with historical data. (Refer to the pitfalls mentioned in Section 2.1.4).

Open access plays a crucial role in IRs. BioMed Central 'de-mythicizes' some of the issues associated with open access (Biomed, n.d.). The current myths are also representative of some of the concerns that prevent compliance, e.g., 'open-access threatens scientific integrity ...' (Biomed, n.d.). BioMed Central's contribution places the prejudices of the contributors and operators into a perspective that individuals/organizations must deal with. Other authors (Björk, 2004; Foster & Gibbons, 2005; Johnson, 2002; Johnson, 2004; Mackie, 2004) discuss compliance and acceptance issues as they relate to change in management of repositories. McLaurin-Smith, et al. (2005) discuss how to merge multiple existing repositories into a single repository. This will help in the planning of the eventual merger of the existing proprietary system with the envisaged repository.

It is clear how closely interlinked Informatics and Information Sciences are in the development and planning of an IR. The role of Information Sciences is to identify needs, whereas Informatics has to put systems in place that will meet these needs.

University of Pretoria etd – van der Merwe, A (2008)

Some of the concerns raised by authors are answered in the next section, especially that regarding copyright issues.

2.1.9.1 Pre-publications and other issues regarding copyright materials

The question of preservation and digital curation of the research outputs requires understanding from all the stakeholders. There are two issues involved, namely ownership and copyright. Achievement of the delicate balance between acknowledgement, ownership and incentives presents a challenge. Authors are often reluctant to part with their intellectual property. Any policy addressing these sensitive issues needs to find a balance between two issues that are often in conflict with each other: namely visibility and preservation of IP rights.

SHERPA's RoMEO (Publisher's copyright and archiving policies) project (SHERPA, 2007) provides IR staff with an explanation of the existing policies of individual publishers in terms of the pre- and post-print versions of articles published in specific journals. There are four categories of publishers' archiving policies, grouped by a colour code.

Table 2: RoMEO Colours (SHERPA, 2007)

| ROMEIO colour | Archiving policy |
|------------------------|---|
| green | can archive pre-print and post-print |
| blue | can archive post-print (i.e. final draft post-refereeing) |
| yellow | can archive pre-print (i.e. pre-refereeing) |
| white | archiving not formally supported |

As Probets and Jenkins (2006) point out, the complexity of Intellectual Property Rights (IPR) is of great concern to all involved with the management and accessibility of intellectual property. It is also not clear what the right approach is in terms of negotiating for IPR with individual publishers and governing organizations. Clarity is also required to determine for exactly how long IPR exists in terms of accessibility via an IR. The Audit Checklist (RLG & NARA, 2005) points out that it is the responsibility of the IR to have a mechanism in place to track and verify the rights and restrictions applicable to a digital item.

Effective rights management is of the utmost importance and that any deviation from this could damage the reputation of an organization. This includes both copyright

laws and contractual agreements. According to Jones et al. (2006), one of the actions that the organization can take to minimize risks is becoming aware of the importance of obtaining distribution rights to publish material online. Institutions should also be aware of the risk of copyright infringement, including plagiarism and the use of databases. In addition, the risk of defamation, liability for provision of inaccurate information, contravention of national and international laws and compliance with data protection regulations must be considered at all times. Another very important, and potentially costly issue, is the accidental or premature disclosure of information, especially in terms of patent applications (Jones, Andres & MacColl, 2006).

2.1.10 Operational issues

There are some basic operational issues that should be considered prior to the implementation of an IR. Included in these are quality control and standardization and auditing. If these two elements are not in place, the IR might be doomed to failure before it is even implemented.

2.1.10.1 Quality control and standardization

Several authors (Anuradha, 2005; Barton & Waters, 2004; Björk, 2004) discuss issues regarding quality and standards, and incentives. One of the tools available to ensure standardization, if only in terms of content, is the use of the metadata schema. The term metadata refers to the description of information on the Internet. It became part of the Internet vocabulary during 1995 with the emergence of the Dublin Core Metadata Initiative (DCMI) Element Set. The purpose of the DCMI schema is to serve as a tool for sharing information between the Internet and traditional libraries (Caplan, 2003). Metadata are intended to help with identifying, describing, and locating information. Although there is still some disagreement regarding the finer details of the definition, it is important to realise that metadata consist of structured information describing an information resource, irrespective of any additional subjective interpretations regarding format and resources with the purpose of facilitating discovery, use and reusability, management and sustainability (Björk, 2004; Caplan, 2003) . There are three distinct metadata typologies, namely descriptive, administrative and structural. The intended use of descriptive metadata is the indexing, discovery and identification of digital resources. Structural metadata are focussed on internal organization of a resource, e.g. chapters, whereas administrative metadata might include technical information regarding the digitization process (Caplan, 2003; DCMI, 2007).

The DCMI has become one of the best-known metadata schemas used by IRs and is described as being ‘... a general-purpose scheme for resource description originally intended to facilitate discovery of information objects on the Web’ (Caplan, 2003:76). The standard is intended to simplify the creation and maintenance of records; to enable standard terminology to be used; to be international in scope and to be extendable. The DCMI element set originally consisted of fifteen data elements. However, three more elements were added later to address needs specific to digital records. These elements are presented in Tables 3 and 4.

Table 3: Dublin Core Metadata elements – Mandatory elements (DCMI, 2007)

| | Identifier | Definition |
|----|-------------------------------|---|
| 1 | Title | A name given to the resource |
| 2 | Creator (if available) | An entity primarily responsible for making the content of the resource |
| 3 | Subject | The topic of the content of the resource |
| 4 | Description | An account of the content of the resource |
| 5 | Date digital | A date of an event in the lifecycle of the resource |
| 6 | Date Original (if applicable) | Creation date of the original resource |
| 7 | Format | The physical or digital manifestation of the resource |
| 8 | Digitization specifications | Technical information about the digitization of the resource, including hardware and software |
| 9 | Resource Identifier | An unambiguous reference to the resource within a given context |
| 10 | Rights management | Information about rights held in and over the resource |

Table 4: Dublin Core Metadata elements – Optional elements (DCMI, 2007)

| | Identifier | Definition |
|---|--------------------------|--|
| 1 | Publisher | An entity responsible for making the resource available |
| 2 | Contributor | An entity responsible for making contributions to the content of the resource |
| 3 | Type | The nature or genre of the content of the resource |
| 4 | Source | A reference to a resource from which the present resource is derived |
| 5 | Language | The language of the intellectual content of the resource |
| 6 | Relation | A reference to a related resource |
| 7 | Coverage | The extent or scope of the content of the resource |
| 8 | Contributing institution | Formerly known as the ‘Holding Institution’. A consistent reference to the institutional units that contributed to the creation, management, description and/or dissemination of the digital resource. |

The Western States Digital Standards Group (DCMI, 2007) guidelines include a couple of brief directives regarding the capturing of data. These mainly focus on the use of abbreviations, punctuations, capitalization, initial articles [grammar] and character encoding. Character encoding is especially a problem in South Africa, with this country’s large number of indigenous languages and, as such, it is essential to ensure that diacritics are displayed correctly and retrieved effectively. Additional

University of Pretoria etd – van der Merwe, A (2008)

qualifiers can be used to provide refined information regarding a specific resource. One such example is the use of alternative titles, where such a refinement will include any form of the title used as a substitute (DCMI, 2007). Input guidelines are also supplied, referring by example to the way in which multiple creators, subject codes and descriptors should be captured.

Although DCMI does provide a standard to be followed, quality remains a challenge and problematic to enforce. One of the tools available is regular auditing of the repository.

2.1.10.2 Auditing

The RLG (Research Libraries Group) and NARA (National Archives and Records Administrators) have published a draft audit checklist for the certification of trusted digital repositories (RLG & NARA, 2005). By providing examples and explanations, criteria that a trustworthy repository should meet, are provided. The Audit tool itself is provided in Section 3 of the RLG/NARA document (RLG & NARA, 2005). In addition, the document is aimed at assisting the people responsible for repositories and who will be responsible for carrying out the audits and certification process. The audit process itself is divided into the following 21 categories:

- Organizations;
- Governance and organizational viability;
- Organizational structure and staffing;
- Procedural accountability and policy framework;
- Financial sustainability;
- Contracts, licences, and liabilities;
- Repository functions, processes and procedures;
- Ingest/acquisition of content;
- Archival storage: management of archived information;
- Preservation planning, migration and other strategies;
- Data management;
- Access management;
- Designated community and the usability of information;
- Documentation;
- Descriptive metadata appropriate to the designated community;
- Use and usability;

University of Pretoria etd – van der Merwe, A (2008)

- Verification of understandability;
- Technologies and technical infrastructure;
- System infrastructure;
- Appropriate technologies, and
- Security (RLG & NARA, 2005).

Deciding on the most important controls to be incorporated is a subjective exercise. However, accessibility, and legal issues such as contracts and licences, workflow, general security, preservation and data management, could have a dramatic impact on an IR and should receive priority. The issues discussed earlier, e.g. technological and operational issues, will cover some of the other auditing elements. However, the availability of a detailed checklist will enable the repository management team to verify that all elements have been addressed in one way or another.

Some internal control measures are built into repository software such as DSpace. The implementation of these control measures will be discussed in more detail in Section 3.2.2 and is directly influenced by the IR's policy. Internal controls can manage access on various levels, as can be seen in Table 5.

Table 5: Authorization in DSpace (Tansley et al., 2007)

| Object | Authorization | Allowed actions |
|------------|--------------------------|--|
| Community | Add/Remove | User can add or remove collections, sub-communities and/or communities |
| Collection | Add/Remove | Add/submit or remove/withdraw items |
| | Default item read | Inherited read of all submitted items |
| | Default bitstream read | Inherited read of bitstreams off all submitted items |
| | Collection administrator | Edit items or withdraw/expunge items in the collection; Mapping of other items to the collection |
| Item | Add/Remove | Add or remove bundles |
| | Read | Item metadata is always viewable |
| | Write | Modification/editing of an item |
| Bundle | Add/Remove | Add or remove bitstreams |
| Bitstreams | Read | View the bitstream |
| | Write | Modify the bitstream |

The systems developers should ensure the effective incorporation of the internal controls. Without this input, the system will not be able to function effectively.

2.2 Identification of the Gap between Problem Statement and Literature

As the move toward the establishment of IRs is still in its early stages and is mostly driven by academic institutions, there are still some gaps in the literature, especially in terms of non-academic institutions. Although there is a reasonable similarity in the information requirements of academia and research organizations, these environments differ significantly in terms of their clientele, although their legal obligations and financial benefits are similar. To date, very little information is available that focuses on non-academic research organizations such as the CSIR. It will therefore be necessary to adapt available information within the context of a contract driven research organization. The CSIR is subjected to legal obligations applicable to its private clients but also has a huge component of publicly funded research that is subject to another set of rules (CSIR, 2007a). It is not clear how a SET organization can meet the demands of its stakeholders, the scientific community and still honour its legal obligations (Republic of South Africa, 2000). In addition, if the work of scientists/researchers is to be recognised, some type of visibility is required. One form of recognition is the accreditation that they receive following publication of their work. Thus, the ability to publish in an accredited journal outweighs any benefit currently associated with an IR.

Accreditation is an important aspect of publishing, both for the academic world and the SET community. The potential impact of IRs on individual accreditation and how this could be managed are issues that still need to be resolved. However, on a more practical level, it is not clear how the issue of accreditation can be managed and applied outside the formal publishing arena. In general, the research funding that a researcher will receive depends on the individual's accreditation level. The visibility and quality of research are mentioned as one of the benefits of an IR but no indication is given of how this will actually be measured, what the benchmarks are, or of how objectivity can be achieved. Although Björk (2004) makes very specific mention of the problem, he does not provide the solution, except to point out that a change in the evaluation system will be required. The question is therefore why scientists/researchers should submit their work to a repository if they do not visibly obtain any benefit from doing so. Although some accredited journals do allow self-archiving, this does not answer the question raised. The solution to this problem might necessitate a merger between Informatics and Information Sciences.

By comparison with an academic institution in which compliance is optional, it is not clear how it should best be approached in a global SET environment. A SET

University of Pretoria etd – van der Merwe, A (2008)

organization seems to be in a better position to enforce compliance, by using a 'carrot and stick' method. Academic institutions seem to be more dependent on the goodwill and voluntary support of the members of academia than do other organizations. In an organization such as the CSIR, the performance evaluation of an individual researcher is linked to a set of agreed-on Key Result Areas (KRA) that are linked to Key Performance Indicators (KPI). As certified list of published items is one of the measurements used, compliance is in the best interests of the individual. What is not clear is how this can be managed and how values (accreditation) should be attributed to publications within the global SET environment. Some sort of standardised approach is required.

No reliable information could be found in terms of the Return on Investment (ROI) of IRs specifically, as intangible benefits have to be measured. It is also not clear what the 'hidden' advantages of an IR could be, e.g. new research requested as a direct result of work published earlier and how this can be measured. In terms of sustainability, it is essential that these two issues be measured in some way or other. Although statistics regarding use are available, there are no known measurements available in terms of value, increase in income and additional funding opportunities. A way of adding tools to measure the value of the content and the value of the research will not only contribute towards accreditation and therefore compliance, but will also help to measure the ROI of an IR, especially in light of the costs mentioned in Section 2.1.1.3.

What is also not very clearly defined in the literature is exactly what is meant by 'long-term preservation' and 'the life cycle of scholarly communication'. Although technical problems with preservation were discussed, 'long-term preservation' remains undefined. It is not clear what is meant by 'long-term' and whether or not 'into perpetuity' is achievable or even realistic. The debate regarding the 'life cycle' of scholarly communication is also ongoing and to date no consensus has yet been reached. It is not clear how the life cycle of information in terms of day-to-day value and the role of information in terms of the organizational memory can be balanced. The organization memory has largely been ignored in terms of IRs and the focus has been more on the cost of technology than on the value of the content. In any SET environment, especially in terms of knowledge management, the organizational memory must be addressed and the sustainability of the digital format of the IR determined.

2.3 Conclusion

There are still many uncertainties and assumptions associated with the implementation of an IR, its real benefits, the best technology and its structure. It is clear that an IR is based on the character and needs of the organization wishing to implement an IR. Furthermore, compliance remains a problem and active advocating of the product is required. The situation with copyright and publishers has to be resolved. The issue of intellectual property rights should be negotiated between scientists and clients. Research outputs following on publicly funded research should be defined and a suitable course of action decided upon.

It might seem that there are more questions than answers. However, waiting for answers is not the right route to take. The world that an IR is preparing for seems to be a reality, as current trends are already moving in that direction. Development and implementation of an IR is the only way of determining whether it is suitable for a specific organization. If visibility regarding the work and expertise of the organization is required, an IR is one of the tools that can be used. Only time will tell if it is the best tool.

The study areas of Informatics and Information Sciences are completely intertwined in terms of development and implementation of an institutional repository. It is neither wise nor logical to separate individual issues. Strictly speaking, Information Sciences can be regarded as the client of Informatics. As the specifications of the repository are set out by Information Sciences, Informatics must ensure that these specifications can be implemented effectively. The close cooperation between these two areas is reflected in the case study discussed in the next chapter.

3 DEVELOPMENT OF A REPOSITORY

3.1 Introduction

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. It developed into a leading SET research and development organization on the African continent. Its mandate, as specified in the Scientific Research Council Act (Act 46 of 1988 as amended by Act 71 of 1990) (Republic of South Africa, 1988), states that: 'The objects of the CSIR are, through directed and particularly multi-disciplinary research and technological innovation, to foster, in the national interest and in fields which in its opinion should receive preference, industrial and scientific development, either by itself or in co-operation with principals from the private or public sectors, and thereby to contribute to the improvement of the quality of life of the people of the Republic, and to perform any other functions that may be assigned to the CSIR by or under this Act.'

About 40% of the annual income of the CSIR is in the form of a parliamentary grant and the balance is generated by research contracts with national, provincial and municipal governments, the private sector and national and international research funding organizations. The core domains within which the researchers operate are biosciences, built environment, defence, peace, safety and security, materials science and manufacturing and natural resources and the environment. The CSIR also explores new areas of research, for example nanotechnology, synthetic biology and mobile autonomous intelligent systems. There are also three national research centres that focus on ICT, laser technology and space-related technology. The researchers are supported by a R&D outcomes portfolio that manages the intellectual property of the CSIR, technology transfer and knowledge dissemination (CSIR, 2007a). The CSIR's Information Services (CSIRIS) forms part of the R&D Core group and assumed responsibility for the development of the CSIR's institutional repository after restructuring of the organization in 2005. The conceptualization and project-planning phase of the IR started during the second half of 2006 and its actual development started in January 2007.

Known as Research Space, the CSIR's institutional repository was launched on 1st August 2007. A copy of the CSIR's home page is shown in Figure 5. Access is provided via the CSIR's homepage, which is shown in Figure 6. The motivating factor for the development of the IR was to increase the awareness of the work done

University of Pretoria etd – van der Merwe, A (2008)

by the researchers in the organization. In addition, the IR is intended to serve as a custodian tool for the data, information and the explicit knowledge of the CSIR. Because of contractual work and client confidentiality, it was not feasible to open-up existing databases to the international research and scientific community. Because of this additional legal complexity, it was decided to develop an IR that would contain – and be limited to – all the publicly available publications of the CSIR and thereby therefore form a subset of the restricted database.

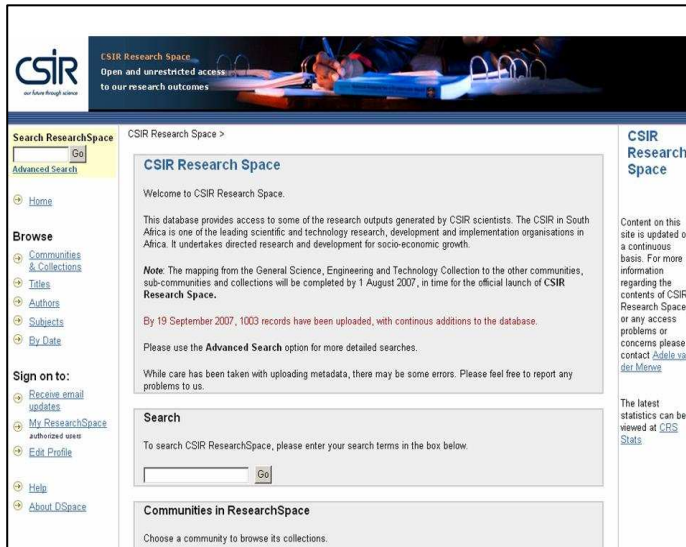


Figure 5: Research Space homepage (CSIR, 2007b)

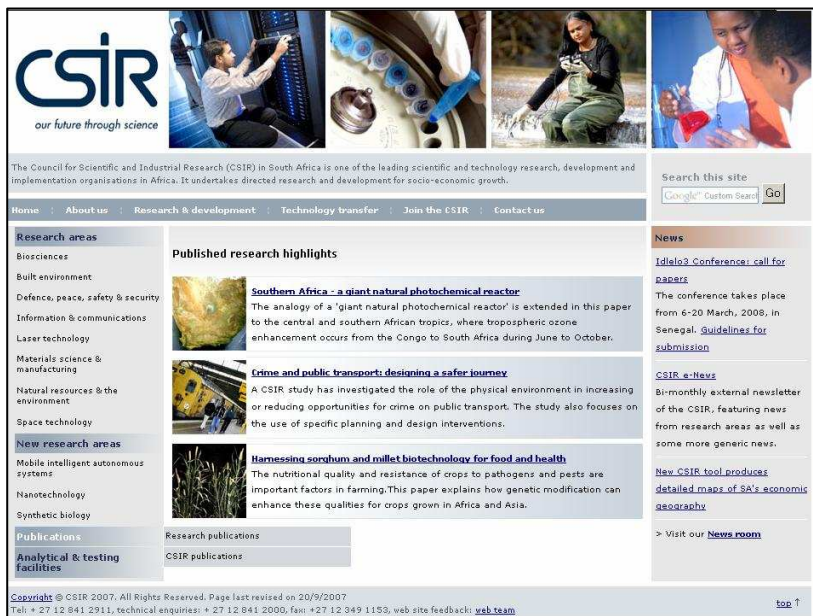


Figure 6: CSIR Homepage (CSIR, 2007a)

University of Pretoria etd – van der Merwe, A (2008)

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 CSIRIS, the CSIR's Computing Services (ICT), Communications, R&D Outcomes and EBAS (Enterprise Based Applications and Systems) groups. During the project planning phase it was decided that DSpace would be the most suitable tool, as this would be able to interact with existing systems, as well as with systems to be planned later, especially in terms of document management and archiving. In addition, the ICT group had already tested DSpace and had started with a trial document-archiving project. Preconceived and negative perceptions were avoided, as most of the people concerned were not acquainted with the product. Knowledge and ideas were shared freely during the design and development of the IR. The use of DSpace is also in line with the move of the organization towards an OSS environment. The CSIR is driving the OSS initiative in South Africa and formally adopted OA in October 2006 as part of its refocused vision.

Hardware issues were also resolved and a dedicated file server was purchased. The University of Pretoria's Academic Information Services (UP/AIS), having already developed and implemented an IR (Smith, 2007c), supported the team with advice and constructive criticism. UP/AIS also hosted the CSIR's trial repository on their file server from January 2007 until July 2007. Their assistance, in the true spirit of OSS, is highly appreciated and it accelerated the whole development process.

With the software and hardware issues having been resolved, it took approximately five months of dedicated teamwork 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 available. Although the information 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 c) to ensure that the data were also reflected in the 'mother' database (TOdB). In the spirit of the IR principles, free and open access to the content of the IR to all interested parties, nationally and internationally, is provided.

Currently, the contents of the IR cover all articles, peer-reviewed, non-peer reviewed articles, conference papers and conference presentations. Also included are other publications of the CSIR, e.g. annual reports, CSIR E-news, as well as the journal, CSIR ScienceScope. A collection of mining related reports has also been added to

the repository. Investigations are currently underway to include multi-media materials such as videos and audio files, although the size of video files is a concern in terms of over-loading the existing bandwidth. Compression technology is being investigated. Alternatives are being sought to solve this problem and to prevent overloading of the existing bandwidth. The team is currently investigating how the different supporting datasets will be managed. This investigation includes identification of the most effective manner to link supporting datasets with specific publications.

3.2 User Requirements and Specifications (URS)

The first challenge that had to be addressed was that of finding the most suitable technology. The decision made will be discussed and justified in more detail later in this section. Table 1 in Section 2.1.5.1 lists the functionalities that are required by the end-user and is based on the ideal situation. As mentioned, DSpace (DSpace Foundation, 2007) was compared with an existing proprietary product, namely InMagic (InMagic, n.d.). Although InMagic complies with users' demands, it does not meet the organizational policy criteria of being OSS. It would have been necessary to purchase an additional licence, as the existing licence is limited to a specific fileserver. This would also entail payment of annual maintenance fees in order to ensure regular receipt of updates and software patches. The decision was therefore made to waive certain features and functionalities in order to continue with the move towards an OSS environment.

The decision to use DSpace as a platform was also based on using a product that could be customised to meet the final and ideal URS. In most cases, as proprietary systems only make provision for limited customization, OSS had to be used in order to ensure complete compatibility between the various systems in use, e.g. the Oracle-based workflow system and other planned systems, such as the document archival system. In order to meet the immediate demand of providing open access to published materials, the shortcomings of DSpace were deemed to be acceptable. Issues not listed in Table 1 related to user acceptance and ease of use. However, these issues are largely based on individual perceptions and customization can address most of the 'non-user-friendly' issues. However, from the outset DSpace seemed to meet both these criteria and users' statistics internationally confirmed the acceptance (See Figure 2). As DSpace provides free and open web-based access, passwords and user identifications are limited to those users who prefer to make use of the alerting service embedded in DSpace.

It is, however, necessary to point out that the above is based on the assumption that an experienced Java programmer will be able to develop the additional features that are required to meet all the requirements regarding compatibility and ease of use. It is debatable whether or not such detailed and intense development will be affordable in terms of both cost and time, which elements are essential and which are just 'nice-to-have' enhancements of an otherwise suitable product. The experiences of colleagues at other institutions have shown that, as not all modifications to the 'vanilla version' of DSpace transfer successfully to later versions, modifications will have to be repeated after each upgrade (Smith, 2007b). Furthermore, additional developments will have to be prioritized and separate specifications and motivations for changes will be required according to the existing service level agreement with the ICT and EBAS groups. It is essential that developments elsewhere be continually monitored in order to ensure that the optimal benefits are reaped.

However, it is understood and accepted by all parties that CSIR Research Space is a work in progress and that – for the short-term at least – additional development will be required. These modifications were budgeted for in the next financial year, namely 2008/9. The following points are issues surrounding the auditing process, discussed in Section 2.1.10.2, which have an impact on quality and administrative issues:

- The number of validation lists is limited. These need to be maintained by hand but the assumption is that they can be modified to meet the minimum requirements, although their maintenance is a bit more complex than would be ideal. However, the workflow procedure discussed in Section 3.4.1 addresses this issue as well.
- Currently there is not the option to make any field unique. The result is additional labour-intensive work, as each new record has to be checked manually against the system, prior to its being added, as this is the only way to avoid duplication at this stage.
- Log sheets are not readily available, which causes problems when human error occurs. Identifying and fixing problems, e.g. accidental withdrawal, is time consuming.
- In general, the search capabilities of DSpace are elementary. In the SET environment, knowledge workers are used to fine-tuning their searches in the finest detail. The lack of the required search functionalities, e.g. Boolean

searching, has already manifested itself as a problem in terms of high-level information retrieval.

Until these issues are addressed by customisation during the next financial year, the operators of CSIR Research Space will be heavily dependent on the quality of the data in the 'mother' system. However, search capabilities, avoidance of duplicate records (unique fields) and improved log sheets are crucial elements scheduled for improvement. However, the issue of whether additional funding will be made available for these improvements is still under discussion.

3.2.1 Structure and features of repository

The structure of Research Space is based on the research structure of the organization. Figure 7 provides a snapshot of two of the main research groups and their specific research areas. As mentioned in Section 3.1, one of the research domains is BioSciences and another is Built Environment. For each of these two domains, communities within Research Space have been created. These communities were then sub-divided into collections, linking to specific research areas within each domain. The same approach was followed with the other research domains. A detailed breakdown of the structure is presented in Attachment C.

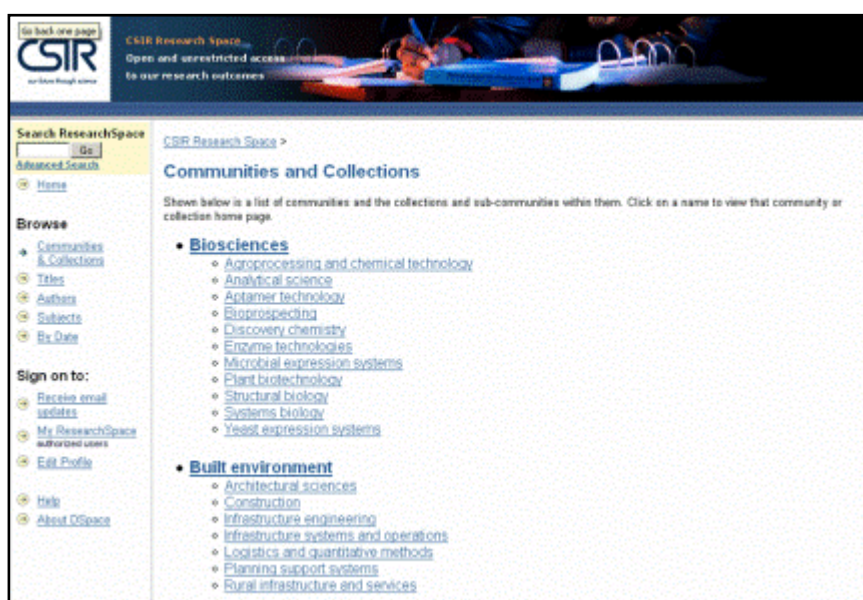


Figure 7: CSIR Research Space structure

However, research organizations such as the CSIR operate in a rapidly changing environment. DSpace can readily accommodate these rapid changes. In order to stay in line with these rapid changes, it was decided to create a 'black bag' in which

University of Pretoria etd – van der Merwe, A (2008)

all the records are stored. Although this was not originally planned, export of the data from UPspace to CSIR Research Space fortuitously resulted in a change in approach. This approach is enriched by mapping from the individual collections within the communities to the holistic collection in order to 'populate' the communities and collections. An additional motivation for the use of a black bag is to ensure that the data are stored in such a manner that access will always be possible and with the least disruption possible, irrespective of changes within the CSIR.

As mentioned earlier, the organization goes through rapid changes. Units can be dissolved as the need for research in a specific area diminishes and new research areas emerge. However, the intellectual property of the organization must still be managed and made accessible long after the original research was done, not only because it is part of the organization memory but also because it is important to safeguard for future reference information generated by the organization. The items stored in the 'black bag' are therefore recorded permanently and will only be withdrawn if any copyright violation, confidentiality or IPR infringement is identified. The decision was made not to delete any items but rather to suppress items until they can be released. As DSpace does not have a readily accessible log, it is necessary to keep a manual list of all items suppressed. However, it is anticipated that the workflow mentioned in Section 3.4.1 will facilitate the record keeping of these items.

Editing access to the 'black bag' is strictly controlled. The 'black bag' approach helps to ensure that items are not deleted by mistake should a unit be dissolved, or a collection deemed to be 'outdated'. In future, communities and collections can be removed or moved without the fear of data loss or corruption. New communities and collections can be added without any concern that the structure may grow out of proportion, thus making retrieval cumbersome. It was not logical to categorize all items into a single collection, as the organization regularly works across silos. By using the mapping approach, a web of interaction can be created without cluttering the system, as mapping is used effectively to ensure comprehensive categorization of information.

Items of a more general nature are also entered into the 'black bag' but separate communities have been created to address this need, e.g. 'CSIR Publications' that covers general research items published by the organization and which discusses the work of the CSIR. Another community is the 'General research interest'

University of Pretoria etd – van der Merwe, A (2008)

community that hosts items applicable to general/holistic perspectives of science, engineering and technology as a concept. The communities and collections approach also enables the user to browse and identify the research areas currently applicable to the organization.

The screen layout of the system provides intuitive access to the information contained in the repository. As such, it meets the requirement of a user-friendly approach. A hyperlink on the left hand side of the screen provides the users with access to the complete structure of the repository. As mentioned earlier, the structure of the IR was based on the structure of the organization. Although some communities have at present only one collection, the structure allows for the inclusion of additional collections and even of sub-collections, should the need arise. Currently, ad-hoc requests for additional communities and collections are received. Examples of this are request to add a Community entitled 'CSIR research featured in mass media' and the request of one of the CSIR's units to create a collection to which all the research done for a specific government department can be mapped. Provided that such requests comply with CSIR policy, additional communities and collections will be added.

3.2.2 *Internal controls*

DSpace provides the functionality to grant access according to a community, a collection, an item, a bundle or a specific bitstream. In Section 2.1.10.1 the internal controls of DSpace were discussed very briefly. By removal of any of the options mentioned, the associated authorization can be removed or blocked. Authorizations can only be done by means of changing the applicable policies on a community, collection or item level. Examples of screen dumps are provided in Figures 8 & 9 below. In terms of Research Space, a predefined group, the Submitter group, has the sole responsibility of adding new items to the repository.

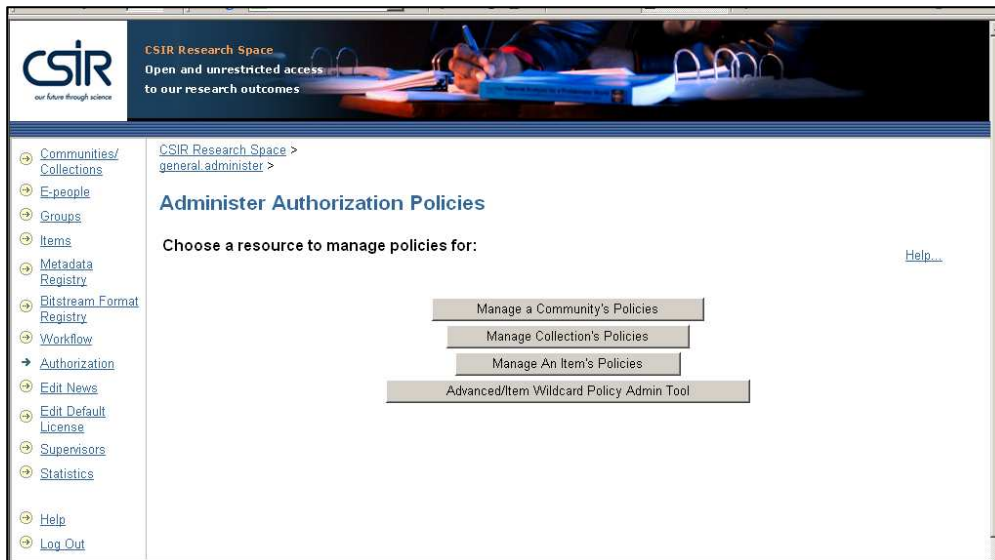


Figure 8: Authorization management by means of policies

Changes to the DSpace policies are illustrated in Figure 8 and impact on the access rights provided. Within each community, the following changes are possible:

- The community's global authorizations can be modified. These modifications will be applicable to every collection and item within the community. Currently, view rights have been given to the Anonymous group, editing rights to the Submitters group (See Figure 9) and administrative rights, including authorisation for inclusion, to the Administrators group.
- Within a community, a specific collection's policies can be modified. These modifications will only be applicable to that specific collection and the sub-collections associated with that collection. Modifications will thus not have any impact on the rest of the collections or sub-collections within that or any other community. This provides the ability to fine-tune policies while minimizing corruption of the system. However, in Research Space this is set to the default setting of the community, as it was not deemed feasible to change any rights on this level.
- A specific item's access can be modified, e.g. only the bitstream can be suppressed or the item itself can be suppressed until further notice. Once again modifications are item specific and therefore do not affect the rest of the items in any of the communities, collections or sub-collections. As in the case of collections, the policies associated with an item default to the policy linked to the community.

It is therefore clear that the administrator of the IR has the ability to adapt specific needs, requirements and limitations from a specific/global level right down to the files attached to a specific entry/item.

Policies for Collection "CSIR e-News" (hdl:10204/1213, DB ID 167) [Help...](#)

Add New Policy

| ID | Action | Group | | |
|-------|--------|-----------------------|------|--------|
| 52761 | ADD | COLLECTION_167_SUBMIT | Edit | Delete |

| ID | Action | Group | | |
|-------|------------------------|-----------|------|--------|
| 52760 | DEFAULT_BITSTREAM_READ | Anonymous | Edit | Delete |

| ID | Action | Group | | |
|-------|-------------------|-----------|------|--------|
| 52759 | DEFAULT_ITEM_READ | Anonymous | Edit | Delete |

| ID | Action | Group | | |
|-------|--------|-----------|------|--------|
| 52758 | READ | Anonymous | Edit | Delete |

Figure 9: Access rights according to groups

As can be seen in Figure 9, final approval of submissions is apparently lacking. This is because of the document-management-workflow process and the fact that CSIR Research Space is a subset of TOdB (See Section 3.4.1 for more information). The quality control of the indexing and authorizations required has already been monitored and verified by the time the item is included in CSIR Research Space. However, it is important to note that quality control still takes place although in an ad-hoc manner. Spot checks are done to ensure that the items included were a) authorised for inclusion; b) contain all the elements required; and c) to ensure that the correct bitstream is added.

3.3 Policies, Procedures and Managerial Reports

The second challenge listed was the development and implementation of policies. Reasonable success was obtained in meeting this challenge. Access to information is important in increasing the visibility of researchers and the organization. It is therefore essential that the information be indexed, archived and preserved with due diligence. The repository was developed and implemented to meet this demand and now needs to be managed and populated in a way that fulfils the expectations of the stakeholders (Van der Merwe, 2007).

CSIR's policy seeks to address the issues surrounding the selection criteria, indexing and archiving of CSIR research outputs, irrespective of the final format. The policy therefore has implications in terms of the responsibility of the organization to

University of Pretoria etd – van der Merwe, A (2008)

show due diligence in obtaining copyright clearance, acknowledging and protecting contractual obligation, as well as intellectual property rights. Although the existence of a repository is not subject to any legal requirements, it is directly linked to the organization's Key Performance Index (KPI) drive and career ladder policy. National and international copyright laws, as well as South Africa's Right of Access to Information Act (Republic of South Africa, 2000) directly influence the repository. Additionally, the repository is influenced by the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (Max Planck Society, 2003) and the Budapest Open Access Initiative (BOAI, 2007) in terms of functionality and services.

The content submission policy is applicable to all appointed research staff members on all the physical sites, buildings, and offices where the CSIR has a research personnel presence. All contractors, sub-contractors, consultants and research service providers while in the service of the CSIR are also included. Lastly, all management and support services personnel also have to submit their output.

The policy is managed and administered by the CSIR Research Space administrator and will be updated as and when required. The policy can be summarized as follows (A copy of the complete draft policy is available as Attachment A):

- Free and open access is provided.
- All formal externally published materials, with the proviso that the required copyright clearance is obtained, are included. CSIRIS is responsible for obtaining copyright clearance as required and for keeping a record of the permissions received.
- Peer-reviewed publications will get preference.
- Research Space will form a subset of the CSIR's Technical Outputs Database (TOdB).
- Authors do not retain personal copyright.
- Information will be managed and curated in accordance with existing standards and policies.
- Only items generated by CSIR personnel will be included.

As it was not deemed feasible to include all historical publications at the outset, a management decision was made to include only items from 1990 onwards, as well as older items with a proven record of demand. In the case of historical publications, digitization will take place on demand and only full text items will be included. In general, obtaining permission for historical items will not be a problem, as only those

University of Pretoria etd – van der Merwe, A (2008)

items of which the CSIR is the sole copyright owner will be considered for inclusion. Some more recent contractual reports, where clients have given their written approval, will also be included, but only on demand.

In terms of usage and the behaviour required by all stakeholders, the policy also includes the metadata, data, submission and preservation policies as subsets of a more comprehensive policy. The complete Metadata policy is available as Attachment B.

Note: The policies are currently only available as draft versions. All the essential issues were however addressed. The policies will be submitted for final approval during 2008.

3.4 Compliance

The third challenge that had to be faced was that of compliance and quality. As with most databases, accurate usage must be monitored and certain administrative information is required to determine the sustainability and ROI of the system. Managerial reports are therefore important, especially when measuring impact scores and the *h-index*. Currently, managerial reports are done manually with data obtained from the system's statistics. The *h-index* was developed by the physicist Jorge Hirsch and is designed to distinguish influential scientists from those who proliferates, e.g. quality (cumulative impact and relevance) vs. quantity (Hirsch, 2005). The *h-index* is relatively effective for the comparison of researchers working in the same scientific field calculation and monitoring of this is labour intensive. Researches across fields cannot be compared with each other because of the different cultures in the various research fields. However, this information is required in terms of funding opportunities and career advancement. It is therefore essential that a method be devised whereby the items accessed via the IR are evaluated in the same manner as used by ISI Web of Knowledge (Thomson Scientific, 2007) and Scopus (Elsevier BV, 2007). This is a problem experienced within the international scientific community for which a solution still has to be found. This shortcoming did not, however, influence the decision to make use of DSpace, as none of the readily available products provides this option.

As a reputable SET organization, the CSIR is particularly concerned about the contravention of any of the legal implications mentioned in Section 2.1.9.1. For that

University of Pretoria etd – van der Merwe, A (2008)

reason, it was agreed not to include contractual items and patents pending until all legal issues had been resolved. Where possible, the organization should retain the intellectual property rights at the time of creation. It should be noted that the rights do not refer back to the individual author/researcher. However, it is important to note that some of the work done by CSIR knowledge workers is done under the auspices of an academic institution while obtaining a postgraduate qualification. In such cases, the academic institution has the right to archive the item and negotiations will then be required prior to the inclusion of an item on CSIR Research Space. Negotiations will also be undertaken with the publishers and external copyright owners to reach some type of middle way that will satisfy all parties, thereby not only giving the author the recognition deserved but also giving the various institutions involved, the exposure and recognition that they deserve. The same is applicable to joint ventures where clarity and agreement is required in terms of making information available.

In terms of protecting the existing IPR, empowered individuals within each unit, e.g. the IP rights manager, will identify those items suitable for inclusion in CSIR Research Space. It is essential that authors have the assurance that both their own and the organization's IP rights will be respected and protected, as well as those of their clients. These issues are addressed in the policy relating to the repository. The workflow procedure discussed in the Section 3.4.1 below is intended to resolve most of the compliance issues.

3.4.1 *Procedures and processes*

The sourcing of information is different from those normally associated with an IR, as self-archiving by the author is neither supported nor encouraged at this stage. As mentioned earlier, the IR is a subset of the TODB system, which is used to calculate publication equivalency information and contains managerial data. The logic behind this approach is that TODB is subjected to strict access and quality control issues. As TODB is located behind the firewalls of the organization, contains classified information, and does not provide access to the actual full text, it is not feasible to provide the wider scientific community access to the database. It is also important to ensure client confidentiality and adherence to contractual obligations. However, submission of information in TODB is obligatory and the information is used to calculate the individual author's KPIs. Lack of compliance will therefore negatively influence the individual researcher. TODB compliance is therefore high and this enables the CSIR Research Space team to source reliable information on a timely

University of Pretoria etd – van der Merwe, A (2008)

manner. As CSIR Research Space is a subset of TODB, the repository also reaps the benefit of the planned workflow process.

The workflow process forms part of a global document management process that will be implemented within the organization in March 2008. The workflow is being developed jointly by CSIRIS and the EBAS group. Figure 10 is a graphical representation of the wider workflow process. Some of the valuable elements of the workflow process are:

- Part of the workflow is a step whereby the author(s) will be required to indicate the collections to which the item should be mapped. This not only ensures a high quality of categorization but also allows the author to provide input before any items are submitted.
- Approval for inclusion in CSIR Research Space can be granted when inclusion is at the discretion of the IP manager. This approval will be monitored and recorded. The Research Space manager is flagged whenever such a request is activated.
- The authors are kept up-to-date with the progress being made and they are informed at the completion of the process. The authors can monitor the status of their publications at any given time.
- Finally, the author will be required to approve the quality of the metadata and to request changes if so required. Once signed off, the data is deemed to be correct and of an acceptable quality.

Figure 10 illustrates the management of public available items, e.g. journal articles and conference papers. Similar procedures were designed for other types of publications, such as contract reports and technical reports. It is anticipated that a major change management process will be involved when the workflow procedure is launched. Some resistance to this is expected and it will be necessary to communicate the benefits to the authors. Every effort has been made to ensure that the process is streamlined. Unfortunately no additional information is yet available regarding the acceptance or rejection of the workflow. Technical problems have delayed the launch of the workflow test phase.

The identification of items for inclusion in CSIRIS Research Space is a very important functionality of the workflow process, as a 'blanket approval approach' is not suitable for the organization. One important difference between TODB and CSIR Research Space is that TODB contains comprehensive bibliographical records of all items, including those of classified items not deemed suitable for CSIR Research Space. It is therefore the responsibility of the CSIR Research Space team to verify eligibility prior to the inclusion of any items in CSIR Research Space, as stipulated in the policy provided. CSIR Research Space staff will be notified by the workflow system of the existence of items for inclusion. An author cannot contact the CSIR Research Space staff and request inclusion of his work unless the inclusion is authorised by the workflow process. Once it has been established that the item is eligible, copyright approval is confirmed or requested from the client or copyright owner. This rather rigid approach is intended to solve existing non-compliance issues.

It is the responsibility of the CSIR Research Space team to manage copyright issues and to ensure that a record is kept of all authorisations obtained. Archiving of the item on Research Space only takes place after all the administrative issues have been completed. However, if it is not possible to obtain the required copyright clearance, the author will be informed and the item will be flagged on the workflow system as not suitable for inclusion. Figure 11 provides an 'expanded' cutout of the procedure as it directly affects the repository and adheres to the draft IR policy. After an item has been indexed and mapped to the applicable collection(s), the URI of the item is entered into the workflow system and the author(s) are notified. This provides the authors with a final opportunity to ensure that they are satisfied with the quality and accuracy of the work done. Final archiving will then take place and the item will be available for general and open access.

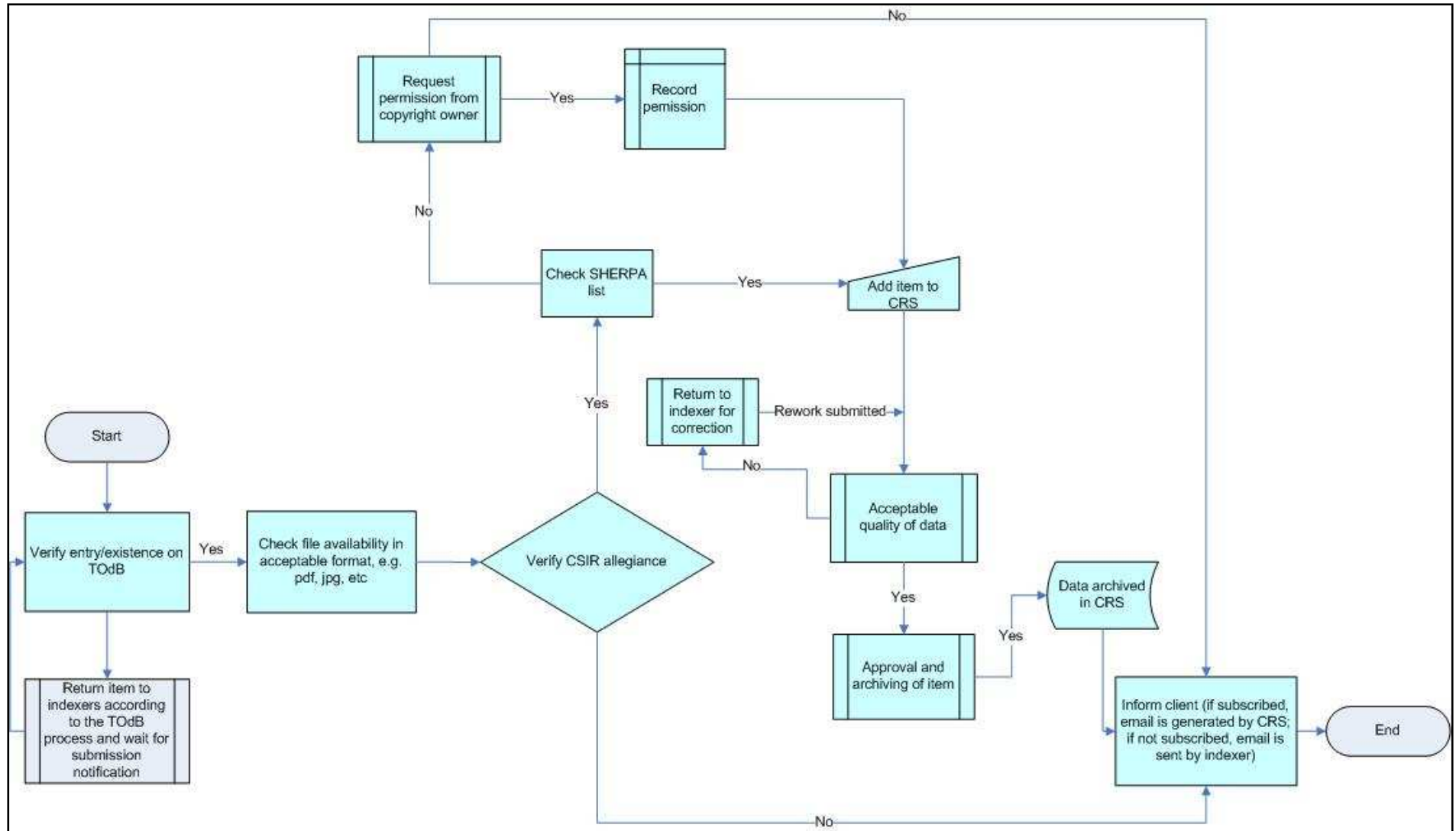


Figure 11: CSIR Research Space workflow

3.4.2 *Standards and quality control of metadata*

Part of the challenge to improving and ensuring compliance is the challenge of improving quality by applying standards. One of the drawbacks of DSpace is the lack of an embedded quality control system. This is one of the reasons for deciding that the CSIR Repository would form a subset of the TODB system as TODB has embedded quality control systems in place. One of the greatest quality problems is the format of the author's name, e.g. surname plus all initials, surname plus nickname or surname plus first name. Although DSpace prompts for surname and initials and provides the user with examples, the system does not provide the indexer with a reliable validation list of authors to make a selection. Although it is possible to develop such a list, it is time consuming, labour intensive and prone to errors because of the human element. The indexer is not always acquainted with all the details regarding individual authors and their allegiance to the organization. Unless some prior knowledge exists, it is very easy to make a mistake at this point.

The TODB system identifies the author according to his/her staff number. The staff number is linked to the HR system and the author's name is derived from that, e.g. surname plus all the initials, including the authors' affiliation with a specific unit. Historical data is available on the HR system although obtaining this information is a bit more complex. However, if the staff number is available, historical allegiance can be traced. The ability to trace the history of the author enables the CSIR Research Space indexer to select the most suitable community and collection, especially in terms of historical information if the author is no longer an employee of the organization. By using the staff number as the primary entry point, it is possible to link all the items written by a specific author and to use the standard format of the entries. It is therefore logical to harvest the information from TODB, a system that is already subject to stringent quality control measures in order to ensure that the quality of the data in CSIR Research Space are of an acceptable level. It also eliminates the need for additional development within DSpace to accommodate the quality requirements of the organization.

In addition, a group of experienced indexers is used to index the items captured in the TODB system. Supporting documentation is available to all indexers, irrespective of experience, e.g. the guidelines prepared by Van der Merwe (2005). CSIR Research Space indexers are not very highly qualified, as the organization makes use of recently graduated interns. As part of their in-service training and in line with

the organization's commitment to skills development within South Africa, these young people are given the opportunity to develop specialised skills. Although they are able to harvest the majority of the items from the TODB system, they still need to add specific information, e.g. the citation. The interns are also taught not to accept any item at face value but to critically evaluate the entry and to correct any errors that they identify, e.g. spelling mistakes, incorrect usage of singular and plural for keywords. In CSIR Research Space, a full and accurate citation must be supplied. As it is essential that this is done according to the Harvard reference style, the interns are provided with a supporting document prepared by Van der Merwe (2006). As the indexers are unable to harvest this information from TODB, they must use their own skills and initiatives, especially in terms of non-standard items, e.g. presentations given to clients vs. presentations given at a conference. The Dublin Core Metadata (DCMI, 2007) also provides the indexer with the guidelines required to complete the data capturing. The specifications and definition of each field are discussed and very little room for interpretation errors remains.

The functionality, usage, and value of any IR are highly dependent on the quality of the contents, i.e. the indexing. Quality monitoring and control thus remain important issues regarding the IR. The fact that quality assurance is split between two systems is irrelevant at present, as the final product and not the process should be judged. As mentioned earlier, the repository is a work in progress and all burning issues will be resolved during customization.

3.5 Determining success

The repository was officially launched on 1 August 2007. However, the system was online and has been available since 15 June 2007. The decision was made to test the IR prior to its formal launch. The reasoning for this was that the project team had to determine the impact on the bandwidth as the content and usage of the repository grew and to take corrective action if needed. It was also necessary to determine whether the original decision to use DSpace was sound and workable. Finally, the project team had to identify any problems and address the problems as soon as possible and to take corrective action, prior to the official launch.

The site at that time displayed a disclaimer indicating that the repository was still under development. The expected launch date was also announced. It is emphasised that the development referred to here is in terms of the actual

information system itself. Despite some glitches in terms of the availability of staff members, the project team was able to meet the target date of 1 August 2007. The statistics generated by DSpace software show a dramatic increase in usage, which has exceeded all expectations. The statistics used reflect the status as it was on 3rd December 2007.

3.5.1 *Visibility and Usage Statistics*

During the first four 'official' months of the IR's existence, figures indicate dramatic and satisfactory growth in the number of both 'once-off' visitors and repeat visitors, as can be seen in Table 6. The pattern is valuable, as detailed information regarding access via unique IP (Internet Protocol) number since the implementation of the IR is not available. Although the growth among repeat visits is not as high as the unique visits, the fact that there is a growth is very important. In November 2007, the percentage of new visits decreased dramatically. As there is a limit in terms of target audience numbers, this drop was expected. However, the consistent increase in repeat visits is extremely gratifying, as it shows that the IR is providing a valuable service to its target audience as users keep coming back for more. On the other hand, unique visitors can be ascribed to a variety of reasons, for example, curiosity to see what is included and browsing to determine the quality of the data. The embedded statistics of DSpace, the functionality expanded on by the ICT team, provide valuable insight into the usage of the system.

Table 6: Usage August 2007 - November 2007

| Month | Unique visitors | Estimated repeat visits | Total number of visits | Pages | Hits | Band-width (GB) |
|-------|-----------------|-------------------------|------------------------|-------|--------|-----------------|
| Aug | 1972 | 1219 | 3191 | 48718 | 78924 | 1.53 |
| Sep | 4731 | 2453 | 7184 | 63074 | 92703 | 4.45 |
| Oct | 8243 | 4217 | 12463 | 83147 | 129179 | 6.77 |
| Nov | 11795 | 5071 | 16866 | 89194 | 100335 | 10.75 |

Compared to the averages listed in the international survey, the usage of the repository can still improve. The mean number of annual visitors in developing countries for one quarter is estimated at 42183 (Primary Research Group, 2007) and Research Space had a very satisfactory 94% (a total of 39704) of that in its first quarter. The data in Table 7 provide a value addition in this regard. It is necessary to separate internal usage from external (outside the organization) usage. In terms of

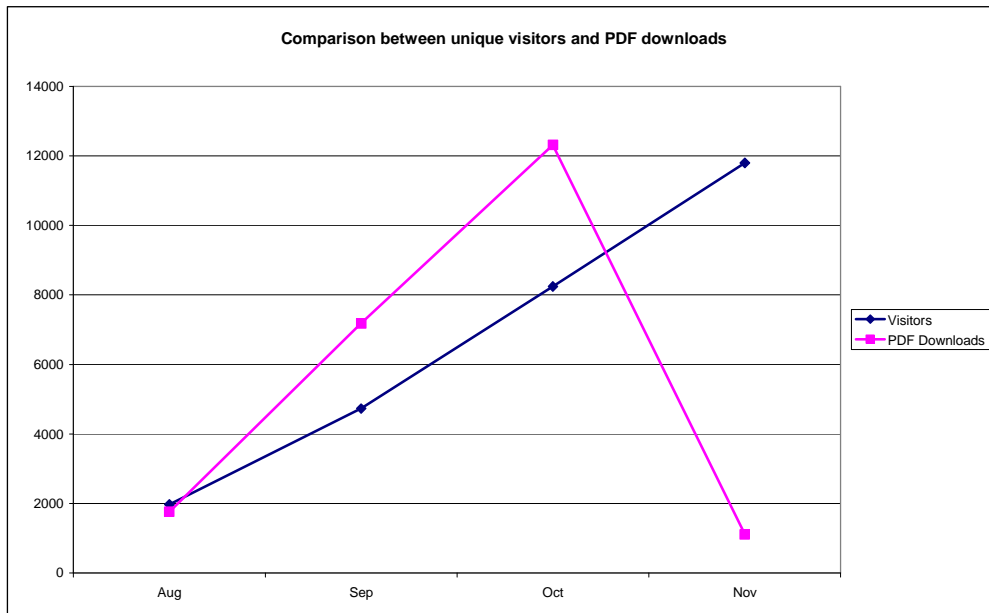
the global picture, internal usage plays a minimal role, as it forms less than one per cent of the overall usage.

One possible explanation of the growth in usage may be the wider visibility of the IR. There is an increase in the number of robots/spiders and search engines harvesting the IR. The number of robots/spiders nearly doubled and the number of search engines is nearly five times higher than in the first month. The number of non-CSIR referring sites more than tripled during the first four months.

Table 7: Monthly breakdown for August-November 2007

| | Aug 2007 | Sep 2007 | Oct 2007 | Nov 2007 |
|---|----------|----------|----------|----------|
| Articles downloaded (pdf files only) | 1756 | 7182 | 12311 | 1111 |
| Harvested via individual robots/spiders | 22 | 38 | 40 | 42 |
| Harvested via individual search engines | 5 | 11 | 20 | 24 |
| Non-CSIR referring sites | 42 | 50 | 90 | 147 |
| Access by non-CSIR individuals | 1648 | 4527 | 8038 | 11652 |
| Access by CSIR individuals | 341 | 264 | 291 | 283 |
| Average view per item | 50 | 285 | 118 | 233 |
| Highest view per item | 161 | 557 | 389 | 265 |

Figure 12: Graphical representation of usage



A situation that will be monitored closely is the balance between the number of visitors and the number of PDF downloads that occur. As illustrated in Figure 12, after October there was a dramatic decrease in the number of PDF (Portable

Document Format) downloads, although the number of visitors and the bandwidth usage increased (see Table 6). It will be necessary to try to identify the cause of the reduction in the number of downloads. This situation is directly linked to the problem of accreditation that was mentioned in Chapter 2. However, it was assumed at the outset that the market would reach a saturation point but, as long as there are downloads, it may be assumed that the repository is achieving its goal of sharing information. In excess of a thousand PDF files downloaded in a month is a significant number, especially in the light of the size of the repository.

Although a mere four months is an insufficient time to identify any reliable behavioural patterns or to predict a long-term trend, the statistics provide insight in the bandwidth used and the potential impact that it might have on the organization as a whole. Although there is a reduction in the percentage increase each month, the mere fact that there is still an increase is indicative that the IR is being harvested by more search engines and dedicated services.

The wide distribution of usage in terms of country of origin and domain type is also interesting. Table 8 provides a summary of the distribution. The complete list of countries and the usage is provided in Attachment D.

Table 8: International access - Status in November 2007

| Geographical area | Pages | Hits |
|----------------------------------|-------|-------|
| Africa | 20035 | 34768 |
| Asia | 2061 | 4638 |
| Australasia | 1195 | 2741 |
| Europe and the UK | 9933 | 18665 |
| Middle East | 662 | 1193 |
| North America | 6259 | 8427 |
| South America | 470 | 1075 |
| Other, e.g. islands and networks | 11069 | 22272 |
| Unknown | 51041 | 85791 |

It is very interesting and gratifying to note that global access is taking place. The information provided in Table 8 is indicative that the repository is reaching the global scientific market. It is also gratifying to note that the repository has the greatest usage reported in Africa. This can be inferred as indicating that the African continent

is sharing in data applicable to the continent. As Africa-generated information is difficult to trace, the development of the repository is a step in the right direction.

Insight into the access gained via the various search engines available is provided in Figure 13. Since the launch of the repository, Google has been the leader, with more than 90% of the hits being generated via that search engine. The number of direct links is interesting but is difficult to analyse. In this instance, the originator of the link is not known and the statistic loses some of its value. Still, the majority of access is as a result of direct linking, which can be seen as a positive trend, an indication that inclusion of the repository in searches is regarded as worthwhile.

| Connect to site from | | Pages | Percent | Hits | Percent |
|--|-------------|-------|---------|-------|---------|
| Origin | | | | | |
| Direct address / Bookmarks | | 19917 | 47.5 % | 21965 | 49.3 % |
| Links from a NewsGroup | | | | | |
| Links from an Internet Search Engine - Full list | | | | | |
| - Google | 18190 18190 | 18945 | 45.2 % | 19431 | 43.6 % |
| - Windows Live | 224 224 | | | | |
| - MSN Search | 189 189 | | | | |
| - Google (cache) | 142 628 | | | | |
| - Yahoo! | 66 66 | | | | |
| - AOL | 42 42 | | | | |
| - Unknown search engines | 34 34 | | | | |
| - Dogpile | 8 8 | | | | |
| - AltaVista | 8 8 | | | | |
| - Netscape | 6 6 | | | | |
| - Others | 36 36 | | | | |
| Links from an external page (other web sites except search engines) - Full list | | | | | |
| - http://www.csir.co.za | 509 509 | 2929 | 6.9 % | 3044 | 6.8 % |
| - http://www.csir.co.za/biosciences/bioprocessingandproductdevelop... | 133 133 | | | | |
| - http://www.aardvark.co.za/search/AfricanSearch.php | 103 103 | | | | |
| - http://www.csir.co.za/Built_environment/Construction/index.html | 101 101 | | | | |
| - http://www.csir.co.za/Built_environment/Architectural_sciences/i... | 77 77 | | | | |
| - http://www.csir.co.za/biosciences/Drug_therapeutic_discovery.htm... | 74 74 | | | | |
| - http://www.csir.co.za/nre/pollution_and_waste/index.html | 69 69 | | | | |
| - http://intraweb.csir.co.za/csiris/databases.html | 61 61 | | | | |
| - http://www.csir.co.za/dpss/ss.html | 58 58 | | | | |
| - http://www.csir.co.za/lasers/laser_physics_technology.html | 57 57 | | | | |
| - Others | 1687 1802 | | | | |
| Unknown Origin | | 53 | 0.1 % | 55 | 0.1 % |

Figure 13: Search engine connections - November 2007

Daily visits remained relatively consistent and show the expected drop over weekends. Although the data presented are limited to November 2007, the same pattern was apparent in the other three months. In future, it will be interesting to monitor peak-usage months and how these affect local bandwidth issues – if at all. Figure 14 illustrates the daily usage during November 2007.

| Day | Number of visits | Pages | Hits | Bandwidth |
|-------------|------------------|---------|---------|-----------|
| 01 Nov 2007 | 594 | 2628 | 4653 | 260.10 MB |
| 02 Nov 2007 | 562 | 1776 | 3632 | 263.50 MB |
| 03 Nov 2007 | 344 | 934 | 1783 | 112.35 MB |
| 04 Nov 2007 | 351 | 1068 | 2038 | 144.80 MB |
| 05 Nov 2007 | 719 | 2938 | 5502 | 386.20 MB |
| 06 Nov 2007 | 729 | 2630 | 4612 | 409.31 MB |
| 07 Nov 2007 | 731 | 3715 | 6016 | 535.17 MB |
| 08 Nov 2007 | 751 | 3406 | 5592 | 325.05 MB |
| 09 Nov 2007 | 576 | 5368 | 9447 | 706.00 MB |
| 10 Nov 2007 | 293 | 1768 | 3112 | 244.43 MB |
| 11 Nov 2007 | 330 | 1600 | 3268 | 234.13 MB |
| 12 Nov 2007 | 669 | 7598 | 12772 | 627.58 MB |
| 13 Nov 2007 | 679 | 6658 | 11358 | 687.30 MB |
| 14 Nov 2007 | 636 | 5680 | 9518 | 588.42 MB |
| 15 Nov 2007 | 630 | 7104 | 11416 | 650.36 MB |
| 16 Nov 2007 | 627 | 7174 | 11552 | 626.49 MB |
| 17 Nov 2007 | 288 | 1814 | 3552 | 222.51 MB |
| 18 Nov 2007 | 308 | 1984 | 3780 | 248.02 MB |
| 19 Nov 2007 | 652 | 4209 | 7522 | 441.08 MB |
| 20 Nov 2007 | 646 | 3148 | 5130 | 320.24 MB |
| 21 Nov 2007 | 681 | 3139 | 5514 | 418.18 MB |
| 22 Nov 2007 | 575 | 3700 | 5670 | 246.36 MB |
| 23 Nov 2007 | 501 | 2876 | 5296 | 253.89 MB |
| 24 Nov 2007 | 267 | 707 | 1572 | 108.06 MB |
| 25 Nov 2007 | 326 | 1618 | 3037 | 181.18 MB |
| 26 Nov 2007 | 685 | 2792 | 5139 | 286.52 MB |
| 27 Nov 2007 | 766 | 3155 | 6680 | 322.24 MB |
| 28 Nov 2007 | 691 | 3368 | 5276 | 286.66 MB |
| 29 Nov 2007 | 667 | 2670 | 4440 | 572.60 MB |
| 30 Nov 2007 | 592 | 3110 | 5209 | 301.08 MB |
| Average | 562.20 | 3344.50 | 5802.93 | 366.99 MB |
| Total | 16866 | 100335 | 174088 | 10.75 GB |

Figure 14: Daily visits - November 2007

Table 9: Items with the 12 highest view scores

| Item/Handle | Number of views |
|--|-----------------|
| Crime and public transport: designing a safer journey (Kruger, T et al) (10204/1028) | 1,107 |
| Southern Africa - a giant natural photochemical reactor (Diab, RD et al) (10204/861) | 763 |
| An appetite suppressant from Hoodia species (Van Heerden, FR et al) (10204/765) | 739 |
| Harnessing sorghum and millet biotechnology for food and health (O'Kennedy, MM et al) (10204/1040) | 649 |
| Model of the transverse modes of stable and unstable porro-prism resonators using symmetry considerations (Burger, L et al) (10204/1293) | 545 |
| Implementing logistics strategies in a developing economy (Iftmann, HW et al) (10204/1139) | 493 |
| Towards a semantic web layered architecture (Gerber, AJ et al) (10204/1189) | 443 |
| Influence of temperature, grain size and cobalt content on the hardness of WC-Co alloys (Milman, YV et al) (10204/1490) | 418 |
| Influence of halogen salts on the production of the ochratoxins by aspergillus ochraceus wilh (Stander, MA et al) (10204/1896) | 416 |
| Measurement of vertical motions of bulk carriers navigating in port entrance channels (Moes, J) (10204/1030) | 410 |
| http://researchspace.csir.co.za/handle/10204/1370 | 406 |
| http://researchspace.csir.co.za/handle/10204/1671 | 395 |

The information provided in Table 8 is further enhanced by the data available in Table 9. Within the organization, individual researchers are evaluated on the number of times their research papers were cited by their peers (the *h-index*). Although the repository has not yet been available for a significant period, the statistics provided in Table 9 provide individual managers with insight into the usability of the work done by the individual researcher. It also enables the organization to identify areas for additional research, as well as those areas no longer in demand. A full list of hits can be drawn from the statistics kept by the system, ranging from a single view up to the

largest number of hits on any given day. This is the first time that this type of statistics can be provided easily and inexpensively on an organizational level. However, as can be seen with the last two entries displayed, the format of the data is not consistent and must be rectified. Verification of the data by opening the links will create a false and inflated statistic. One option is to change the entry to reflect a true citation style. The development of this functionality will be investigated during the 2008/2009 financial year.

3.6 Skills development

As an organization, the CSIR is regarded as a national asset and is therefore committed to knowledge sharing, technology transfer and the skills development of South Africans. Based on this approach and on the general lack of existing expertise in terms of IRs, a decision was made to make use of recently graduated students (interns) to assist with the project, especially with the population of the repository.

At the start of the project, the CSIRIS team consisted of the project leader and four interns. The first three months were dedicated to development of the indexing skills of the interns and to introducing them to the DSpace software. The interns were also used during the comparison phase of two potentially feasible systems. During the next three months, the students were empowered to function independently even though their work was still submitted to ad-hoc quality control processes. After the first six months, two of the interns gained long-term employment elsewhere. The remaining two interns were then tasked with ensuring that the repository remained viable, that it contained data of high quality in terms of indexing and with taking care of all the administrative issues, e.g. copyright clearance. By the end of July 2007, one of the interns was asked to assume responsibility for the day-to-day administration of the repository. The other intern was entrusted with starting the preparatory work for another planned repository. However, both interns were required to support each other and to ensure that the data were entered into the system within twenty-four hours after they were received and that the standards were adhered to at all times.

This approach proved to be very successful and very satisfying for all involved. With the assistance of the interns, it was possible to populate the repository with close to 1000 items within the first three months. The success of this approach also resulted in the decision to continue with this approach for the immediate future. The current

interns also assisted with the training and orientation of the new interns who started in January 2008. It is not anticipated that this approach will change in the near future.

In addition, close cooperation with colleagues at the University of Pretoria's Academic Information Services (UP/AIS) yielded invaluable results. Because of this interaction, it was possible to fast track the development. The development and implementation process was reduced by at least fifty per cent. The CSIR project team was able to avoid pitfalls experienced by the UP/AIS team and in return, to offer some solutions to the problems experienced by the UP/AIS team. In general, the CSIRIS team was able to benefit greatly from its interaction with its UP/AIS colleagues.

On the technical side, the assistance of the ICT group proved invaluable. Its expertise, especially in determining the compatibility with the existing organizational infrastructure and plans, enabled the project team to be more focussed. A lack of Java programming skills within the CSIRIS group was counterbalanced by the expertise within the ICT and EBAS groups. The project proved yet again that combination of the right skills from different sectors will result in a successful IS product.

3.7 Conclusion

It is unrealistic and false to pretend or insinuate that the project progressed smoothly and without any problems. Problems did occur, especially in terms of time-management and quality, and corrective action was called for. To insinuate that the success of the project was based solely on the efforts of the CSIRIS project team would also create the wrong impression. The cooperation and support of colleagues and stakeholders contributed greatly towards its success. This emphasizes the important role that external parties can play in the development of any project. Knowledge sharing proved to be extremely valuable.

The success of the project can also be attributed to the [accidental] timing of the project. The time was right, as attitudes and perceptions are slowly changing. The decision to make use of interns to populate the repository proved to be wise, although this originally met with some resistance. Not only were the interns able to dedicate their time and attention to developing the project, they were also eager to develop their personal skills. This led to a "win-win" situation, as the team was able

to populate the repository in record time. In addition, the interns' hard work and dedication enabled the team to start with the second phase much sooner than anticipated.

The statistics proved to be very satisfying. By being able to show the usage of the repository, some of the remaining reservations are being eradicated. Authors realize that they have an important role in the success of the repository and are starting to become pro-active by ensuring that their information is reflected in the repository. A benefit not originally anticipated relates to the use of the statistics in negotiation for additional research funding. Although detailed records of expenditure were not kept (other than of the salaries of the interns and the purchase of the file server), no unexpected expenditure occurred.

This project also emphasized the fact that Informatics play a crucial element in other subject fields such as, in this instance, Information Sciences. To provide a good information service, an excellent information system is required. The two areas share the same passion for sharing information and only through combination of their individual areas of expertise will it be possible to develop an excellent information system. To do so, the two groups will have to continue to work together and to share their knowledge, insights and ideas. Through the willingness of all concerned to listen to colleagues and identification of existing best practices, the project proved to be a success.

The next chapter discusses the results, feedback and discussions that the CSIRIS project team recorded. Information provided in Chapter 4 will help to place some of the statements made in Chapter 3 in context. However, it is pointed out that the project is still in its early stages and that and only time will tell if the current optimism is justified.

4 RESULTS AND DISCUSSION

As part of the marketing 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 promoting a move away from impersonal statistics. Several issues kept on surfacing and were dealt with during the presentations, namely copyright, intellectual property right issues and client confidentiality. The concerns raised by the individuals proved that previous assumptions that these two issues would present a major challenge were correct. The planned implementation of a formal workflow process helped to alleviate most of the concerns. Although all possible precautions are taken to prevent a contravention of the two issues, end-users still need to be convinced that their interests will be protected.

As some authors have already made use of the system, they gave valuable feedback based on their perceptions of the repository. Most of the comments confirmed existing perceptions, especially in terms of the statistics. An interesting comment was made during one presentation, namely that the speaker was using the repository as a marketing tool. The repository enabled him to negotiate successfully for an increase in research funding. The request for a better individual item usage-listing format was raised at most of the meetings. This confirms the opinion of the development team that the present format is inadequate. Another concern was that not all the historical information was included from the beginning. It was necessary to convince the authors that this would be done in a timely manner. Authors were also informed that any errors would be rectified and that any omissions would be monitored against the TODB system. It is important that authors realise that the ultimate responsibility still rests with them.

The implementation of the repository proved to be well timed and very successful. Although some problems and delays were experienced, the project generally went according to plan. Some unexpected crises regarding the availability of staff were absorbed without causing any major problems or delays. This is mostly due to the trial period offered by UP/AIS and the lessons and advice from which the CSIR was able to draw.

The launch of the repository also led to some unexpected events. As the team leader's contact details were displayed on the repository, she started to receive calls from users interested in additional research. Currently there is no structure or

process in place to handle such request/queries. This issue will have to be addressed as soon as possible.

4.1 Lessons learned

When work on the repository first started and, to justify the expenditure, the team was required to compare a legacy system with the new planned system. This resulted in running a dual system and keeping statistics of transactions. As, because of time constraints, it was deemed illogical to build a repository using DSpace software, an alternative had to be found. As mentioned earlier, a solution presented itself in the form of the UP/AIS offer to host the CSIR's trial repository on its system.

The comparison was completed within three months and the decision made to continue with the DSpace project. At that stage, the team was confronted with some unforeseen issues. Exporting and importing of data to and from DSpace are done using Java programming rather than tab-delimited ASCII (American Standard Code for Information Interchange) as was originally anticipated and assumed. The team therefore had to make the following decision: a) Develop CSIR specific structure and re-do all the work/data capturing that had been done to date or b) Use the structure as developed by UP/AIS and adapt this where needed. Because of time constraints it was deemed illogical to re-do three months' work. The decision was made to keep with UP/AIS' structure, because the changes required were mostly of a cosmetic nature. However, during the transfer of the data from UPspace (<https://www.up.ac.za/dspace/>) to the CSIR platform some data were lost. In order to identify the lost items, the records had to be checked manually and this proved to be very labour intensive. The lack of easily accessible logs regarding activities proved to be a problem, which will have to be addressed in the future.

Another lesson learned by the team was that items that were suppressed had to be logged manually for future references. It seems that DSpace does not keep a central record of these activities although a record does exist in the metadata (provenance) of the record itself. It will therefore be necessary to keep a record of the URIs assigned by the system for future use and reference. This has been addressed in the workflow system but the success of the system still needs to be determined.

Working in a virtual team environment proved to be both challenging and frustrating. The different representatives were subjected to various other demands on their time

and skills. Unforeseen crises at times prevented the timely delivery of components, e.g. branding. This shouldn't be considered as a reflection on individual competencies or dedication but rather an acknowledgement of the complexity surrounding virtual teams. The need for concise and clear communication also became apparent as misunderstandings regarding deliveries and responsibilities occurred. This could be attributed to off-line and unmonitored informal meetings at which decisions were made but not recorded.

4.2 Problems experienced

The fact that the project was completed in a record time should be taken into consideration. Delays prior to the actual development of the project resulted in the loss of safety margins and a pragmatic acceptance of what could be done within the allotted time span. Although the deadline of the project was not changed, all non-crucial customisation activities planned were shelved temporarily.

Some of these non-crucial customisation issues are the validation lists and internal audit control that are currently in use within the TOdB system. For example, without specific and additional programming, it is not possible for the system to identify duplicate records or to link the author's name to a staff number. The result was that the identification and removal of duplicate records had to be done manually. Variants in the format of the author's names are also a problem, e.g. the inclusion of all initials vs. the author's first initial only. These are typical problems associated with harvested data. It was decided not to customise DSpace at this time but rather to make use of the workflow system to address this shortcoming.

The workflow is currently under development and is scheduled for launch during March 2008. It makes provision for the identification (flagging) of items. Central to the system is the correct format of the author's name. This should contribute to an improvement in quality and adherence to existing standards. The ability of the author to select the most suitable community and collection applicable to the work will also eliminate errors resulting from a lack of subject expertise.

The original repository structure planned was very simplistic and limited to three communities, namely Science, Engineering and Technology. As the project neared the completion date, the structure had to be revised completely. However, the decision regarding the 'black bag' approach made earlier proved to be beneficial.

When the data were transferred from UPSpace to CSIR Research Space, these were dumped into a single community/collection because of time constraints. At that time, the team was unable to assign items to communities and collections effectively. By storing all the data in a single 'black bag' collection it was possible to painlessly change and modify the structure to suit everybody. As awareness of the repository grew, it became necessary to add new communities and collections. This was done quickly and without any problems. The 'black bag' decision also proved to be beneficial in view of anticipated changes during the lifetime of the repository.

The search functionality of the system is, however, still a concern and will have to be addressed urgently. The system is unable to search for words within a title as the search is based on the first word of the title. It is also not possible to search for a single word within a phrase and therefore the correct keyword should be used. As it is impossible to anticipate all possible phrases, the system will have to be upgraded to search for words within phrases as well as at the beginning of sentences or search phrases.

The support from the ICT group, although very valuable, was limited on account of circumstances beyond its control. These led to misunderstandings and different interpretations. They also caused delays and the project effectively started five months after the original planned date. It became clear that insufficient time had been allowed for the different phases. This resulted in some of the phases, e.g. intensive testing, being ignored in order to make up for lost time. Fortunately, as a result of the prior experience gained while using the UPSpace site, corrective action could be taken prior to implementation, thereby eliminating most of the on-site testing. However, this is not the ideal and might still prove to be a very costly decision.

4.3 Evaluation and reasoning

In general, the project may be regarded as a rousing success, exceeding all expectations. The statistics proved to be more exhaustive than originally thought. In addition, usage of the repository has exceeded all expectations and is growing every month. However, it is anticipated that the growth will level out or even decrease, as the addition of new items to the repository will eventually slow down. Discussions during subsequent road shows confirmed the perceptions that the repository had succeeded in meeting its goals. There are also an increasing number of requests to

increase the number of additional phases, particularly as regards the inclusion of identified historical research reports.

One of the early concerns was that usage of the system could result in over-utilisation of the available bandwidth. This fear was partially allayed by the fact that no problems that could be connected with the repository were reported or identified. The file server is stable and the system is online well within any reasonable limits. Since the launch of the system in August 2007, some small problems regarding the availability of statistics were experienced but these were speedily addressed. However, something of greater concern is a seeming instability regarding the rights of the different groups. It was necessary to reset some individual item view rights of the anonymous group. In an attempt to identify the source of the problem the occurrence of this problem, as well as any error messages is now being monitored.

On the technological level, the standards of the repository's design and content meet international standards. The repository was successfully registered with a variety of harvesters including OAIster and Scopus. Usage of the repository is also evident in the ratings of Google when applicable items are retrieved. Test runs done prior to these registrations of Research Space showed that the repository rating moved from the tenth to the first position, in terms of relevancy.

4.4 Recommendations

The development of the IR proved yet again that nothing proceeds according to plan and that the project team had to be ready, willing and able to adapt to changing circumstances. The value of networking and sharing of knowledge proved to be invaluable in ensuring that the project was completed within the allotted time span. The team's flexibility, with the required level of quality and expertise being retained, proved to be one of its most valuable features. The project was completed according to specifications, within budget and on time and the results were more positive than originally anticipated.

However, mistakes were made. It is necessary to acknowledge these and to list the actions taken to address and resolve the resulting problems.

The first problem occurred during the trial phase and resulted from the decisions taken at the time. It is not feasible to use a temporary structure with the aim of

fine-tuning it later. DSpace is not flexible enough to accommodate fine-tuning at a later stage, as all existing records have to be changed manually. This will result in a lot of repeat work and will prove to be costly unless an alternative solution is found. The project team resolved this issue by changing its approach – a decision which, although proving to be beneficial, could easily have resulted in abandonment of the project.

The second problem occurred during the planning of the structure. The original decision was to keep it simple and to move away from the structure of the organization. This decision proved to be unacceptable to the stakeholders and had to be adapted during the development phase. A lack of consultation between the team leader with all the stakeholders to identify their needs was the direct cause. Adaptability again enabled the project team to resolve the matter quickly and easily.

The lack of standards and quality control systems within DSpace proved to be a bigger problem than anticipated. However, implementation of the workflow system will address this problem and it will also be of value should DSpace in future be replaced by another system. Also, as the repository is a subset of another quality-controlled system, this shortcoming can be accommodated. It is not clear what will happen when the ‘mother-system’ changes or how this will affect the repository. Any changes to either system will have to consider the impact on the other.

Compression of video files and long-term preservation of all formats are still challenges that must be resolved. Although plans for these have been tabled, they have not yet been tested and this is still a major concern in terms of sustainability. The whole issue regarding obsolete or outdated software and hardware must be investigated properly. Risk assessments must be done and projects launched to formalize the migration of data to current products. This is an ongoing process and to date there has been a lack of sufficient or ongoing planning.

The following recommendations should be of value to anybody planning an IR.

- Verify the accuracy of statements and avoid ‘technical’ assumptions. For example, the IR team assumed that the export of data from UPSpace could easily be manipulated and moved to other communities and collections. This proved not to be the case and an alternative had to be implemented. In general, this can be attributed to assuming that shared understanding existed. This not only caused delays at the beginning of the project that could have been very

costly at the end, but also necessitated changes to the original project plan. Especially when people from different backgrounds are involved, it is essential to ensure that there are no misconceptions and to communicate any changes as soon as possible. Prior experience with similar products created a false sense of security that might have proved to be costly.

- Communication proved to be a challenge. When a dedicated team is not available, communication increases in importance and it is essential that the team leader be kept aware of what is happening in the other areas, as well as where there were necessary changes in priorities as a result of unforeseen circumstances. As the project team consisted of members from other units, it was often difficult to get the entire team together on short notice. This resulted in having to rely on email communications for decision-making, leading to miscommunication and delays. The result was conflict and stress that could have been avoided.
- All issues regarding branding should be resolved prior to the implementation of the system. Because of communication breakdowns, the branding of the repository was delayed. Although this did not influence the developmental work, it could have delayed the launch of the official product.
- The structure of the repository should be finalised prior to implementation of a trial project. Working as close to reality as possible is the only reliable way of testing the product. It might not be easy or even possible to 'fix things' with the final product.
- Proper planning should resolve many issues and would make the process easier. However, time should still be allocated to resolve issues such as staff turnover, essential expansion of the original scope of the project and delays caused resulting from the unavailability of critical personnel. It is therefore essential to plan for backups and to be open to alternative solutions.
- As the work was done by a virtual team, typical problems, e.g. breakdown in communication and delays due to external influences had an effect on the smooth running of the project. However, as the virtual team was geographically situated on the CSIR campus, it was able to solve problems quickly and effectively. However, should the members of the virtual team be geographically separated, care should be taken to arrange video conferencing for effective resolution of problems.

4.5 Expected Contribution to Knowledge

There is a plethora of information on institutional repositories as it relates to the academic world but very little is available in terms of a SET organization. The role and value that a repository has within the scientific research community needs to be quantified and perhaps even justified. It is the intention to bring the problems and potential solutions facing the SET community to the fore.

The South African Research community is faced with the same challenges that any other research community faces. On the one hand, the need for recognition and acknowledgement for researchers is just as important - if not even more important - as for academics. Recognition determines their future research funding and determines their self-worth, just as it does with academic researchers. Nevertheless, the major obstacle that the researcher faces repeatedly is the issue of confidentiality. Researchers in a research-based organization are often confronted with a situation in which they need to publish their work but are prevented from doing so owing to contractual constraints. However, unless an article is published in an acknowledged peer-reviewed journal, it is impossible to receive the acknowledgement and accreditation sought. Furthermore, an increased demand by end-users has also shown that research artefacts should preferably be published in an open-access domain. An effectively developed and managed repository as a facilitation tool could help to alleviate these challenges. Although the repository itself will not resolve the issues mentioned, managed accesses can serve as a step in the right direction.

The contents of this work are aimed at decision-makers, librarians, archivists, document managers and other stakeholders within the scientific community who are faced with implementing repositories for their organizations. The insights gained in finding workable solutions and making recommendations should assist other organizations to focus their activities and, it is hoped, to share their activities with the scientific community. The intention is to fill the existing gap regarding information about SET institutional repositories. It is also the intention to improve understanding of the contractual obligations that often prevent researchers within the SET environment from sharing the results of their research. This dissertation provides a workable compromise between the legal obligation to provide information and the client's right to privacy.

4.6 Future research

There are two potential research areas, namely the use of IRs to determine *h-index* and determination of the ROI of the repository, thereby ensuring the sustainability of the repository.

4.6.1 *Determining the h-index*

Calculation of an author's *h-index* is very complex and labour intensive. Existing repository software does not address this issue. In order for a SET and academic institution to determine the value an author's work the repository should be developed to measure and calculate the *h-index*. Research on how to implement this will be required. Systems will have to integrate with other international and national databases to log the number of times a particular item has been cited, the publications in which it was cited and whether or not it was cited in a positive manner. A way must be developed whereby works by single and multiple authors can be evaluated in such a way that they reflect the value of an author's work.

4.6.2 *Determining the Return-on-Investment*

Developing and maintaining a repository is costly, as mentioned earlier. In order to ensure sustainability it is essential for the organization to determine its return on investment. Currently this is not being done in terms of IRs. Issues that will have to be addressed are the potential increase in funds, the increase in research opportunities, the value of the repository in terms of visibility and acknowledgement, the measurement of the quality of the research and so forth. The development of a framework of what will be measured needs to be developed and values and weights must be linked to each of the elements. OSS must be used to develop a tool whereby the data can be obtained while removing any subjectivity on the side of the user. At present, the criteria that need to be used and the weight of each are still vague and unexplored. In-depth research will be required to determine how to do this, to identify what needs to be measured and to develop a system that will provide valuable and reliable information.

4.7 Conclusion

The lessons learned and the manner in which the virtual team functioned provided valuable insights into what can go wrong and presented possible solutions that can be implemented. The case study also highlighted the dangers of making assumptions just to save time. Mention is made in the literature of the perceived benefits of an IR for an organization. What is not clear is how these benefits can be measured in terms of ROI and accreditation for individual authors. Additional research is called for to enable workable solutions to be implemented.

5 CONCLUSION

It has been argued that an IR provides a valuable and essential 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 a sustainable endeavour (Anuradha, 2005; Crow, 2002; Johnson, 2002; Lynch, 2003). This study set out to test the claims and to determine whether a service that was developed for the academic sector has a right of existence for SET organizations.

If the definitions of Rankin (2005) and Crow (2002) are taken at face value, it seems logical that an institutional repository is just as valid for a SET organization as it is for an academic institution. Allard et al. (2005), Anuradha (2005), and Lynch and Lippincott (2005) are just some of the authors who attempt to prove that implementation of institutional repositories is a value-added service that can be provided to the scientific and research community. If the short-term results are any indication, an IR should prove valuable in the medium to long term as well, although this still needs to be determined and proven. The short-term statistics of the CSIR Research Space Repository can be regarded as indicating value and are in line with the claims made by several authors, such as Allard et al (2005), Anuradha (2005), Crow (2002) and Rankin (2005). Informal feedback received from authors and managers during the road shows proved that they view the repository in a positive light and that they are eager to give their support towards the long-term availability and sustainability of the system.

A concern highlighted in the literature is the issue of long-term preservation (Bullock, 1999; Harmsen, 2008; Hockx-Yu, 2006; Stanescu, 2005; Wheatley, 2004). Preservation of digital format is more complex than that of paper-based information, mainly due to the rapid advances in technology. Enabling effective usage of digital formats requires detailed planning and budgeting. It is also essential that detailed risk assessments be carried out to determine the effect of changes in media, upgrades, new software and hardware. Care should be taken to ensure that digital formats do not end up in 'digital waste lands' of inaccessible, valueless artefacts. This is especially important if it is considered that the data or information can only increase in value when used repeatedly, that is, when information becomes knowledge. It is essential that preservation must address the issue of perpetuity. Preservation is also further complicated by what are referred to as "intellectual

straightjackets”, e.g. copyright laws, obsolete systems in terms of hardware and software and any other action or inaction that might affect the preservation of data, information and the formats in which they are available. Lack of effective preservation will nullify all other efforts in the curation and accessibility of data and information. Digital preservation does not happen per chance but must be planned for when the service is being developed. At the 1st African Digital Curation Conference held in Pretoria, Harmsen (2008) pointed out that a reliable and trustworthy service, be it an archive or repository, must have suitable preservation policies in place in order to ensure sustainability. Sustainability should therefore not just be limited to ongoing growth (in terms of contents and system development) and financial viability but should also include the long-term accessibility of the information.

The literature (Barton & Walker, 2003; Barton & Waters, 2004; Crow, 2002; Lynch, 2003; Mackie, 2004; Pinfield, 2002) provided valuable guidelines and information regarding the development of the repository. It is essential that changes in technology and new trends be monitored on a continual basis. Since repositories as an open-source and open-access service, have only been available for a very short period, changes are inevitable. It is the responsibility of the repository manager to implement only those changes that will be of value and improve the functionality and features of the repository. For a repository to be sustainable, it is essential that the stakeholders can place their trust in an efficient, well planned and well managed IS system. By executing regular and detailed risk assessments regarding the repository the manager can take reasonable preventative action to ensure that the repository is worthy of trust (McHugh, 2005; McHugh et al., 2007; RLG & NARA, 2005). Issues such as financial sustainability, long-term preservation of digital formats and legal issues such as copyright and intellectual property rights are just some of the issues that require regular risk assessment in order to identify weak spots in the planning, development and implementation of a repository.

The question of whether or not a repository is suitable for a SET organization was answered, albeit more in practice than in theory. The lack of documented proof is attributable solely to the fact that IRs are still in their infancy. However, Anuradha (2005) is of the opinion that IRs have an extremely important role to play for any SET organization. The results following the launch of the CSIR Research Space also proved beyond any reasonable doubt that any SET or research organization needs a repository as a window to its work. The positive reaction of researchers and of the

wider scientific community indicates that implementation of repositories was long overdue. Although some issues still need to be resolved, e.g. preservation and migration of data, the basic workflow is functioning smoothly. There is an awareness of the repository at all levels of stakeholders, including the end-users on both national and international levels. However, what the long-term impact will be in terms of accreditation, acknowledgement, and individual *h-indexes* (Hirsch, 2005) is not yet clear. Also it is not yet possible to calculate the Return-on-Investment as it relates to IRs specifically, mainly because of the complexity of measuring intangible benefits. Unfortunately, preservations, storage and labour costs can escalate. Only time will tell whether IRs can deliver in terms of trust, expectations and long-term sustainability.

The efforts involved in finalising the policies and obtaining the support from the stakeholders should not be underestimated. Although there are very valuable publications available regarding these issues (Allard, Mack & Feltner-Rechert, 2005; Crow, 2002; Foster & Gibbons, 2005; Jenkins, Breakstone & Hixson, 2005; Lambert, Matthews & Jones, 2005) the individual organizational cultures can and do complicate the finalisation of the policies. Working in a rapidly changing environment further complicates the finer details of the policies. Policies should also guide rather than dictate if co-operation and compliance is required. Policies should also be clear regarding the benefits of the service and should include the vision and mission of the service (Barton & Walker, 2003; Barton & Waters, 2004; Crow, 2002; Lynch, 2003; McHugh et al., 2007; Smith, 2006). In addition, policies should also address operational issues such as preservation, quality control, standardization, auditing and obtaining copyright clearance. The roles and responsibilities of the individuals should be defined and communicated clearly. Should any deviation from assigned roles occur without prior approval or notification, stakeholders may lose their trust in the repository. Regular auditing and risk assessment exercises can help to identify potential problem areas and assist with the development of long-term trust and in improving the sustainability of the repository (Barton & Waters, 2004; Lynch, 2003; McHugh et al., 2007).

The sources used for this study and the experience gained during the development of the CSIR Research Space repository provided valuable insight in terms of the planning and implementation of such a service. Additionally the interdependency, rather than pure collaboration, between Informatics and Information Science within this context became very clear. Although the roles are unique, a high level of

interdependency is required to ensure success. During the development of 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 good communication all contributed to the ultimate success of the project. The departure of several team members from the service of the organization since the launch of Research Space project in August 2007 did not impact negatively on the project. It is due to effective planning, documentation, and transfer of skills and knowledge. It also became very clear that a service such as an IR could not – and should not – be implemented in isolation and that cooperation and collaboration play an essential, dynamic and ongoing role in ensuring the sustainability of the service. In the spirit of the OSS movement, the Budapest Open Access Initiative (BOAI, 2007) and the Berlin Declaration regarding Open Access (Max Planck Society, 2003) collaboration between institutions proved to be of mutual benefit, especially in terms of developing the IS system.

6 REFERENCES

Allard, S., Mack, T.R. & Feltner-Rechert, M. 2005. "The librarian's role in institutional repositories: a content analysis of the literature". *Reference Services Review*, vol. 33, no. 3. Available from: <<http://www.emeraldinsight.com/0090-7324.htm>>. [Accessed on 25 February 2007].

Anuradha, K.T. 2005. "Design and development of institutional repositories: A case study". *The International Information & Library Review*, vol. 37, no. 3, pp. 169-178. Available from: <<http://www.sciencedirect.com>>. [Accessed on 1 June 2006].

Bailey, C.W., Coombs, K., Emery, J., Mitchell, A., Morris, C., Simons, S. & Wright, R. 2006. *Institutional repositories*. Association of Research Libraries, Washington, DC.

Barton, M.R. & Walker, J.H. 2003. "Building a Business Plan for DSpace, MIT Libraries Digital Institutional Repository". *Journal of Digital Information*, vol. 4, no. 2. Available from: <<http://hdl.handle.net/1721.1/26700>>. [Accessed on 25 May 2007].

Barton, M.R. & Waters, M.M. 2004. *Creating an institutional repository: LEARDIS Workbook*. Cambridge, MA: MIT Libraries Online. Available from: <<http://hdl.handle.net/1721.1/26698>> [Accessed on 2 June 2006].

Beier, G. & Velden, T. 2004. "The eDoc-Server Project: building an institutional repository for the Max Planck Society". *HEP Libraries Webzine*, vol. 9. Available from: <http://bioline.utsc.utoronto.ca/usage/testcd/fulltext/Institutional_Archives/beier_edoc-server_project.pdf>. [Accessed on 25 May 2007].

Billings, M.S. 2005. *Institutional repositories: sabbatical report January 30 - July 8, 2005*. University of Massachusetts, Amherst. Available from: <http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1000&context=marilyn_billings>. [Accessed on 3 May 2007].

Biomed. n.d. "(Mis)Leading open-access myths". *Open Access Now*. Available from: <<http://www.biomedcentral.com/openaccess/inquiry/myths/?myth=all>>. [Accessed on 2 June 2006].

Björk, B. 2004. "Open access to scientific publications - an analysis of the barriers to change?" *Information Research*, vol. 9, no. 2. Available from: <<http://informationr.net/ir/9-2/paper170.html>>. [Accessed on 15 March 2007].

BOAI. 2007. *Budapest Open Access Initiative*. Available from: <<http://www.soros.org/openaccess/read.shtml>>. [Accessed on 19 October 2007].

Boulanger, A. 2005. "Open-source versus proprietary software: Is one more reliable and secure than the other?" *IBM systems journal*, vol. 9, no. 2. Available from: <<http://www.research.ibm.com/journal/sj/442/boulanger.pdf>>. [Accessed on 15 March 2007].

Branin, J. 2003. *Institutional repositories; draft paper for Encyclopedia of Library and Information Science*. Ohio State University Libraries, Columbus, OH. Available from: <<http://hdl.handle.net/1811/441>>. [Accessed on 25 April 2005].

Broeder, D., Auer, E. & Wittenburg, P. 2006. "Unique Resource Identifiers". *Language Archives Newsletter*, no. 8. Available from: <http://www.mpi.nl/LAN/issues/lan_08.pdf>. [Accessed on 17 January 2008].

Bullock, A. 1999. "Preservation of Digital Information: Issues and Current Status". *Network Notes*, no. 60. Available from: <<http://epe.lac-bac.gc.ca/100/202/301/netnotes/netnotes-h/notes60.htm>>. [Accessed on 27 May 2007].

Caplan, P. 2003. *Metadata fundamentals for all libraries*, Chicago, American Library Association.

Crow, R. 2002. *The case for institutional repositories: a SPARC position paper*. SPARC, Washington, DC. Available from: <http://www.arl.org/sparc/bm~doc/ir_final_release_102.pdf>. [Accessed on 25 April 2007].

CSIR. 2007a. *CSIR - Council for Scientific and Industrial Research home page*. Available from: <<http://www.csir.co.za/>>. [Accessed on 20 September 2007].

CSIR. 2007b. *CSIR Research Space*. Available from:
<<http://researchspace.csir.co.za/dspace/>>. [Accessed on 2007/09/20].

CSIRIS. 2007. Research Space project team: Minutes of meetings .

Darwin, C. n.d. *Charles Darwin Quotes - Quotations from the famous naturalist*. Available from: <http://www.darwin-literature.com/l_quotes.html>. [Accessed on 11 March 2007].

DCMI. 2007. *Dublin Core Metadata Initiative (DCMI)* Dublin Core Metadata Initiative. Available from: <<http://dublincore.org/>>. [Accessed on 1 June 2006].

Devakos, R. 2006. "Towards user responsive institutional repositories: a case study". *Library Hi Tech*, vol. 24, no. 2. Available from: <<http://www.emeraldinsight.com/0737-8831.htm>>. [Accessed on 25 April 2007].

DSpace Foundation. 2007. *DSpace*. Available from: <<http://www.dspace.org>>. [Accessed on 1 June 2006].

Elsevier BV. 2007. *Scopus - Basic Search*. Available from: <<http://0-www.scopus.com.innopac.up.ac.za/scopus/home.url>>. [Accessed on 2007/09/21].

EPrints. 2007. *EPrints for digital repositories*. Available from: <<http://www.eprints.org/>>. [Accessed on 1 June 2006].

Foster, N.F. & Gibbons, S. 2005. "Understanding faculty to improve content requirement for institutional repositories". *D-Lib magazine*, vol. 11, no. 1. Available from: <<http://www.dlib.org/dlib/january05/foster/01foster.html>>. [Accessed on 15 March 2007].

Fowler, H.W., Fowler, F.G. & Thompson, D. (eds). 1995, *Concise Oxford Dictionary of Current English*, Clarendon Press, Oxford.

Gibbons, S. 2004. "Establishing an institutional repository", *Library Technology Reports*, July/August. Available from: <<http://www.techsource.ala.org>>. [Accessed on 15 March 2007].

Goh, D.H., Chua, A., Khoo, D.A., Khoo, E.B. & Mak, N., M.W. 2006. "A checklist for evaluating open source digital library software", *Online Information Review*, vol. 30, no. 4, pp. 360-379. Available from: <www.emeraldinsight.com/1468-4527.htm>. [Accessed on: 20 April 2006].

Halland, Y. 2007. Personal communication. Unpublished.

Harmsen, H. 2008. "The final seal of approval: Directives for data producers/researchers, digital consumers and digital archives", *African digital curation conference (1st : 2008 : Pretoria)*. National Research Foundation, Pretoria. Available from: <http://stardata.nrf.ac.za/nadiccc/presentations/harmsen_henk.ppt>. [Accessed on 15 February 2007].

Hirsch, J.E. 2005. "An index to quantify an individual's scientific research output". *Proceedings of the National Academy of Sciences*, vol. 102, no. 46, pp. 16569-16572. Available from: <<http://www.pnas.org/cgi/content/abstract/102/46/16569>>. [Accessed on 21 September 2007].

Hockx-Yu, H. 2006. "Digital preservation in the context of institutional repositories", *Program*, vol. 40, no. 3, pp. 232.

InMagic. n.d. *InMagic website*. Available from: <<http://www.inmagic.com/>> [Accessed on 15 June 2006].

Jenkins, B., Breakstone, E. & Hixson, C. 2005. "Content in, content out: the dual roles of the reference librarian in institutional repositories". *Reference Service Review*, vol. 33, no. 3. Available from: <<http://www.emeraldinsight.com/0090-7324.htm>>. [Accessed on 21 September 2007].

Jihyun, K. 2005. "Finding documents in a digital institutional repository: DSpace and EPrints", *Proceedings of the American Society for Information Science and Technology*, vol. 42, no. 1. Available from: <<http://dx.doi.org/10.1002/meet.1450420173>>. [Accessed on: 10 May 2007].

Johnson, R.K. 2004. "Open-Access: Unlocking the value of scientific research". *Journal of Library Administration*, vol. 42, no. 2. Available from: <http://eprints.rclis.org/archive/00005089/01/OA-Oklahoma_article.pdf>. [Accessed on 3 June 2006].

Johnson, R.K. 2002. "Institutional repositories: partnering with faculty to enhance scholarly communication". *D-Lib Magazine*, vol. 8, no. 11. Available from: <<http://www.dlib.org/dlib/november02/johnson/11johnson.html>>. [Accessed on 1 June 2006].

Jones, R., Andres, T. & MacColl, J. 2006. *Institutional repository*. Chandos Publishing, Oxford.

Kuny, T. 1997. "A digital dark ages? Challenges in the preservation of electronic information". *63rd IFLA Council and General Conference*. Available from: <<http://www.aiim.org/fbia/documents/63kuny1.pdf>> [Accessed on 15 March 2007].

Lambert, S., Matthews, B. & Jones, C. 2005. "Grey literature, institutional repositories and the organizational context", *Proc. 7th International Conference on Grey Literature (GL7)*, GL7 Program and Conference Bureau, Amsterdam.

Lynch, C.A. 2003. *Institutional repositories: essential infrastructure for scholarships in the digital age*. Association of Research Libraries, Washington, DC. Available from: <<http://www.arl.org/resources/pubs/br/br226/br226ir.shtml>>. [Accessed on 15 March 2007].

Lynch, C.A. & Lippincott, J.K. 2005. "Institutional repository deployment in the United States as of early 2005". *D-Lib magazine*, vol. 11, no. 9. Available from: <<http://www.dlib.org/dlib/september05/lynch/09lynch.html>>. [Accessed on 3 June 2006].

Mackie, M. 2004. "Filling institutional repositories: practical strategies from the DAEDALUS project". *Ariadne*, no. 39. Available from: <<http://www.ariadne.ac.uk/issue39/mackie/>>. [Accessed on 1 June 2006].

Mark, T. & Shearer, K. 2006. "Institutional repositories: a review of content recruitment strategies", *World Library and Information Congress: 72nd general conference and council*. IFLA. Available from <http://www.ifla.org/IV/ifla72/papers/155-Mark_Shreaer-en.pdf> [Accessed on 1 July 2007].

Max Planck Society. 2003. *Open Access Conference - Berlin Declaration*. Available from: <<http://oa.mpg.de/openaccess-berlin/berlindeclaration.html>>. [Accessed on 19 October 2007].

McHugh, A. 2005. *Open Source for Digital Curation*. University of Glasgow, Glasgow. Available from: <<http://www.dcc.ac.uk/resource/curation-manual/chapters/open-source/opensource.pdf>> [Accessed on 25 February 2007] .

McHugh, A., Ruusalepp, R., Ross, S. & Hofman, H. 2007. *Digital repository audit method based on risk assessment (DRAMBORA); Version 1.0 (draft)*. Digital Curation Centre and DigitalPreservationEurope. Available from: <<http://www.repositoryaudit.eu/>>. [Accessed on 25 February 2007].

McLaurin-Smith, N., Young, E. & Sullivan, S. 2005. "'Two into one will go": combining two institutional repositories at University of Melbourne", *ETD 2005 Conference*. Available from: <<http://adt.caul.edu.au/etd2005/papers/056Young.pdf>> [Accessed on 1 June 2006].

Ngwenyama, O.K. & Lee, A.S. 1997. "Communication richness in electronic mail: Critical Social Theory and the contextuality of meaning". *MIS Quarterly*, vol. 21, no. 2. Available from: <<http://www.people.vcu.edu/~aslee/ngwlee97.htm>>. [Accessed on 19 June 2006].

Nixon, W. 2003. "DAEDALUS: Initial experiences with EPrints and DSpace at the University of Glasgow", *Ariadne*, no. 37. Available from: <<http://www.ariadne.ac.uk/issue37/nixon/intro.html>>. [Accessed on: 10 May 2007].

Open Society Institute. 2004. *A guide to institutional repository software; 3rd edition* . New York: Open Society Institute (OSI). Available from: <http://www.soros.org/openaccess/pdf/OSI_Guide_to_IR_Software_v3.pdf>. [Accessed on 1 June 2006].

OpenDOAR. 2007a. *OpenDOAR Chart - Content Types in OpenDOAR Repositories - Worldwide*. Available from:

<<http://www.opendoar.org/onechart.php?clD=&ctID=&clID=&IID=&potID=&rSoftWareName=&search=&groupby=ct.ctDefinition&orderby=TallyDESC&charttype=bar&width=600&caption=Content Types in OpenDOAR Repositories - Worldwide>>. [Accessed on 9 May 2007].

OpenDOAR. 2007b. *OpenDOAR Chart - Usage of Open Access Repository Software - Worldwide*. Available from:

<<http://www.opendoar.org/onechart.php?clD=&ctID=&clID=&IID=&potID=&rSoftWareName=&search=&groupby=r.rSoftWareName&orderby=TallyDESC&charttype=pie&width=600&height=300&caption=Usage of Open Access Repository Software - Worldwide>>. [Accessed on 9 May 2007].

Pinfield, S. 2002. "Creating institutional e-print repositories". *Serials: The Journal for the Serials Community*, vol. 15, no. 3. Available from:

<<http://uksg.metapress.com/media/a861qmvuj2uceq7ta2w/contributions/w/a/w/l/wa/wlp1qld3fn10u.pdf>>. [Accessed on 1 June 2006].

Primary Research Group 2007. *International Survey of Institutional Digital Repositories*. The Group, New York, NY.

Probeta, S. & Jenkins, C. 2006. *Documentation for institutional repositories* Loughborough University. Available from:

<<https://magpie.lboro.ac.uk:8443/dspace/bitstream/2134/782/1/lppaper++final.pdf>>. [Accessed on 1 June 2006].

Rankin, J. 2005. *Institutional repositories for the research sector*. National Library of New Zealand, Wellington.

Republic of South Africa. 2000. *Promotion of Access to Information Act*. Cape Town. Available from: <<http://www.info.gov.za/gazette/acts/2000/a2-00.pdf>>. [Accessed on 5 January 2008].

Republic of South Africa. 1988. *Scientific Research Council Act*. Cape Town.

Available from:

<http://www.info.gov.za/docs/legislation_compliance/scientific_research_act.doc>

[Accessed on 5 January 2008] .

RLG & NARA. 2005. *Audit checklist for the certification of trusted digital repositories*, RLG, Mountain View, CA.

Roode, D. & Byrne, E. 2006. *Research methodology: course INF830*. University of Pretoria, Pretoria.

Sanger, L.M. 2006. "The future of free information". *Digital universe journal*, Article 2006-1. Available from:

<http://www.dufoundation.org/downloads/Article_2006_01.pdf>. [Accessed on 2 June 2006].

SHERPA. 2007. *SHERPA RoMEO project*. Available from:

<<http://www.sherpa.ac.uk/>>. [Accessed on 2007/05/28].

Smith, B. 2002. "Preserving tomorrow's memory: preserving digital content for future generations". *Information Services and Use*, vol. 22, pp. 133-139. Available from:

<<http://www.ebsco.com>>. [Accessed on: 27 May 2007].

Smith, I. 2007a. *IRSpace listserv*. Available online: <irspace@kendy.up.ac.za>.

Smith, I. 2007b. *Personal communication*. Unpublished.

Smith, I. 2007c. *UPSpace at the University of Pretoria*. Available from:

<<https://www.up.ac.za/dspace>>. [Accessed on: 27 May 2007].

Smith, I. 2006. *DSpace implementation: policies, procedures and problems*.

Available from: <<https://www.up.ac.za/dspace>>. [Accessed on: 27 May 2007] .

Stanescu, A. 2005. "Assessing the durability of formats in a digital preservation environment; the INFORM methodology". *International Digital Library Perspectives*, vol. 21, no. 1, pp. 61-81. Available from: <<http://www.emeraldinsight.com/1065-075x.htm>>. [Accessed on: 27 May 2007].

Tansley, R., Bass, M., Branschofsky, M., Carpenter, G., McClellan, G. & Stuve, D. 2007. *DSpace System Documentation*. Available from: <http://www.dspace.org/index.php?option=com_content&task=view&id=151&Itemid=116>. [Accessed on 2007/10/12].

Thomson Scientific. 2007. *ISI Web of Knowledge [v3.0]*. Available from: <<http://portal.isiknowledge.com/portal.cgi>>. [Accessed on 2007/09/21].

Van der Merwe, A. 2007. *CSIR Research Space (CRS) policy; draft*. Pretoria: CSIR. Unpublished.

Van der Merwe, A. 2006. *Citing and referencing: basic guidelines*. Unpublished.

Van der Merwe, A. 2005. *Guide for indexers and abstractors: general principles and practices*. Unpublished.

Van Westrienen, G. & Lynch, C.A. 2005. "Academic institutional repositories: deployment status in 13 nations as of mid 2005". *D-Lib magazine*, vol. 11, no. 9, Available from: <<http://www.dlib.org/dlib/september05/westrienen/09westrienen.html>>. [Accessed on 2 June 2006].

Wheatley, P. 2004. *Institutional repositories in the context of digital preservation*. Digital Preservation Coalition, York, UK. Available from: <<http://www.dpconline.org/docs/DPCTWf4word.pdf>> [Accessed on 2 June 2006].

Attachment A: CSIR Research Space Policy: Stakeholders and compliance

CSIR POLICY; draft

| | | | |
|----------------------|----------------------------|-----------------------|------------------|
| Title | CSIR Research Space | | |
| Policy Number | POL R xxxx | Section | |
| Revision | 0.1 | Effective Date | August 2007 |
| Prepared by | A van der Merwe | Signature | |
| Approver | Group Manager: R&D | Signature | Sibongile Pefile |

1 Purpose

To enhance the use of and exposure of CSIR's intellectual output through its institutional repository – CSIR Research Space. Access to information is important in increasing the visibility of researchers and the organization. For it to be acknowledged as a trusted resource of intellectual content it is essential that information contained in the institutional repository is indexed, archived and preserved with due diligence and in accordance with internationally acknowledged open-access standards.

2 What it seeks to address

The policy seeks to address the issues related to the selection criteria for the inclusion of CSIR research output, in all mediums, in Research Space as well as the indexing and archiving of these records.

3 Implications

The policy has implications in terms of the responsibility of the CSIR to show due diligence in terms of obtaining permission from copyright owners and the stakeholders. It relates to the acknowledgement of contractual obligations, the protection of intellectual property and the potential sensitivity of information. It also has implications in terms of selection of the items included in the repository.

4 Benefits

The repository provides access to full text items thereby contributing to the visibility and credibility of the individual researcher and the organization as a whole. In addition, the repository promotes the long-term preservation and management of scholarly information and ensures that information remains accessible irrespective of the software used to create the document or artefact.

5 Regulatory framework

The Copyright Act, the Right of Access to Information Act and contractual obligations influence the inclusion of items.

The repository is also anticipating future developments regarding open access to public funded research as recommended by the OECD.

6 Links to other policies

- Research ethics
- Data curation
- Research outputs

7 Who is to use it and when

This policy is applicable to:

- all research staff members
- all contractors, sub-contractors, consultants and research service providers while serving the CSIR,
- all management and support services, and
- Research Space administrators.

8 Who should be consulted

- CSIR Fellows
- Researchers
- Operational Unit IP managers
- Legal services (when faced with non-routine items)
- Copyright holders

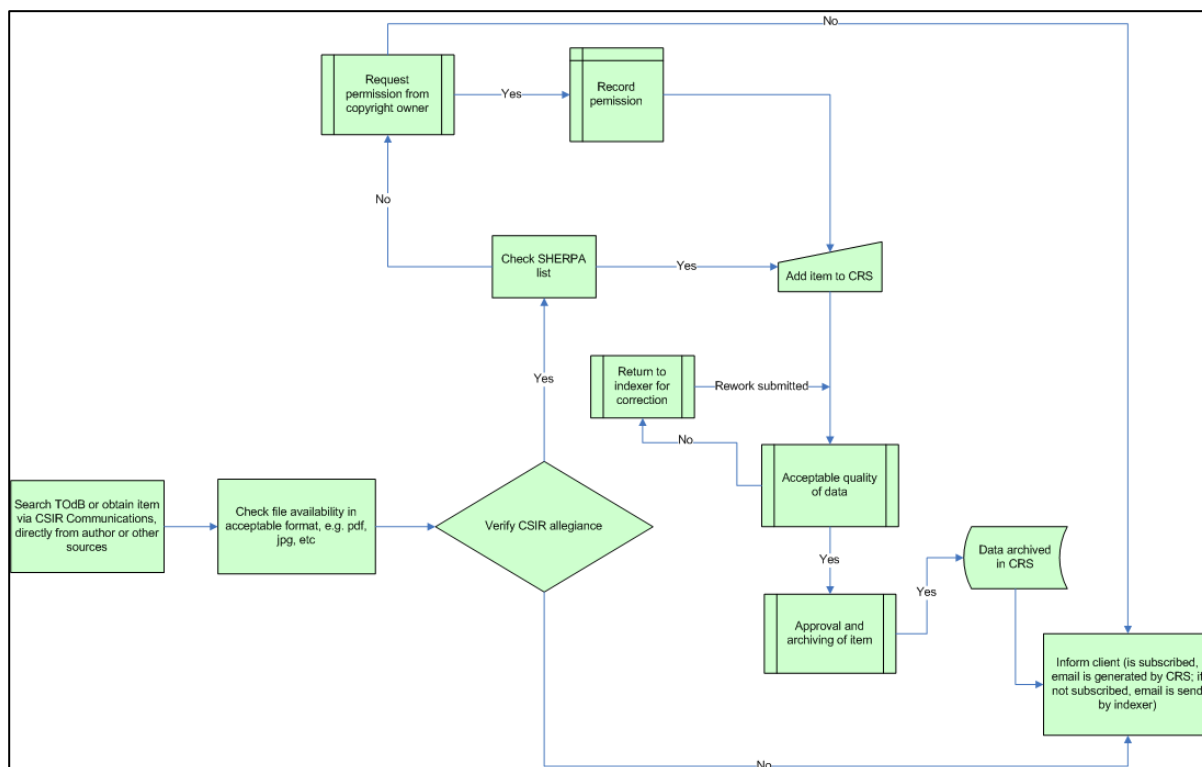
9 Policy statement

The CSIR is committed to providing free and open access to a defined and selected collection of full text research publications, multimedia files such as video and audio items, and datasets associated with or supporting research publications. The necessary infrastructure will be maintained so that the CSIR's repository could retain international status as a trusted repository of high quality research output. To this end the CSIR will:

- Make available all formally externally published materials, with the proviso that the required copyright clearance is obtained, as well as a selection of research reports and other artefacts of research knowledge output.
- Research Space will form a subset of the CSIR's Technical Outputs Database (TOdB) and only publications submitted to the TOdB in electronic format will be considered for inclusion in Research Space. It is the responsibility of the relevant staff member to ensure that research output is submitted to the TOdB.

- Authors do not retain personal copyright, as the copyright rests, with either the publisher or the CSIR.
- Information will be managed in accordance with the records management policy and the applicable data curation standards.
- CSIRIS is responsible for obtaining copyright clearance as and when required.
- All staff will take the necessary precautions to ensure that only approved publications are included in the repository. The inclusion of items not published externally must be approved by the IP Manager of each Operational Unit/Centre to ensure that intellectual property remains protected. **Research reports, manuals, technical notes and legal documents, will only be included in Research Space with written authorization from a member of the Unit's management team.**
- Only items generated while the author is a CSIR employee will be included in the repository.
- Only items where at least one of the authors is a CSIR staff member will be included.
- Theses and dissertations done as part of graduate and postgraduate studies will not be included except in a bibliographical format while providing hypertext links to the academic institution's repository unless copyright clearance is obtained.
- Where available a published document's data set(s) and any other artefacts linked to the document will be made available and accessible via the repository.

10 Process
 10.1 Process flow chart



11 Description of the process

Items are sourced in different ways, namely:

- from the TODB system,
- directly from the author, or
- directly from CSIR Communications.

The availability of the electronic publication is verified according to the SHERPA RoMEO list and the format checked for suitability. IP Authorization or verification for inclusion in the repository is checked. CSIR allegiance and copyright issues are checked. If necessary, copyright approval is obtained and a record (in GWDMS) is kept of the authorization received. The item is indexed, subjected to quality control and on approval, the item is archived and deposited in the repository. The client is informed of any problems experienced along the way and at the end of the process when the item has been archived.

12 Approval

The repository manager checks the quality of the indexing and verifies that the full text item is correct before completing the repository workflow process.

13 Implementation

The policy will be administered by the CSIR Information Services staff with the cooperation of the R&D Manager and Operational unit IP managers.

14 Communication

All research staff must be made aware of this policy document and related standards and procedures. The document must be used during the induction process of new staff members and be easily accessible (e.g. the IntraWeb).

15 Monitoring processes

It is the responsibility of each individual author to ensure that all the publications are included in the TOdB. The author needs to indicate clearly when a record, if ever, may be transferred to the repository. The responsibility of CSIRIS is to ensure that due process is followed in ensuring the legality and quality of the items included.

16 Glossary

Institutional repository: A database providing free and open access to selected full text publication and/or multimedia files.

External publications: Items published in peer-reviewed journals, papers presented at conferences, presentations at conferences and any other item seemed suitable to be made freely available to any interested party.

Data Sets: Any supporting document/record in digital format, e.g. a MS Excel file.

17 Acronyms

CSIRIS: CSIR Information Services

IP: Intellectual Property

Attachment B: CSIR Research Space Policy: Metadata, data, submission and preservation

Note: The policy was developed with the assistance of an online tool provided by OpenDoar. The tool is freely available at <http://www.opendoar.org/tools/policytool.php> and provides an interactive approach for the development of the policy.

- 1 Metadata policy for information describing items in the repository:
 - Anyone may access the metadata free of charge.
 - The metadata may be re-used in any medium without prior permission for not-for-profit purposes, provided:
 - The OAI identifier or a link to the original metadata record is provided
 - CSIR Research Space is mentioned
 - The metadata must not be re-used in any medium for commercial purposes without formal permission.

- 2 Data policy for full-text and other full data items:
 - Anyone may access full items free of charge.
 - Copies of full items generally can be:
 - Reproduced, displayed or performed, and given to third parties in any format or medium
 - Used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided
 - The authors, title and full bibliographic details are given
 - A hyperlink and/or URL are given for the original metadata page
 - The content is not changed in any way
 - Full items must not be sold commercially in any format or medium without formal permission of the copyright holders.
 - The repository is not the publisher – it is merely the online archive.
 - Mention of CSIR Research Space is appreciated but not mandatory.

- 3 Submission policy concerning depositors, quality & copyright
 - Items may only be deposited by accredited members of the institution, or their delegated agents.
 - Authors may only submit their own work for archiving.
 - The administrator only vets items for the eligibility of authors/depositors, relevance to the scope of CSIR Research Space, and valid layout & format.

- The validity and authenticity of the content of submissions is checked by internal subject specialists.
- Items may not be deposited until any publishers' or funders' embargo period has expired.
- If CSIR Research Space receives proof of copyright violation, the relevant item will be removed immediately.
- Once the required approval is received, the relevant item will be restored.

4 Preservation policy

- Items will be retained indefinitely.
- CSIR Research Space will try to ensure continued readability and accessibility.
- Items will be migrated to new file formats where necessary.
- It may not be possible to guarantee the readability of some unusual file formats although all reasonable action will be taken to facilitate the long-term readability of file formats.
- CSIR Research Space is dependent on external partners to back up items in external archives.
- CSIR Research Space regularly backs up its files according to current best practice.
- Items may be removed at the request of the author/copyright holder.
- Acceptable reasons for withdrawal include:
 - Journal publishers' rules
 - Proven copyright violation or plagiarism
 - Legal requirements and proven violations
 - National security
- Withdrawn items are not deleted per se, but are removed from public view.
- Withdrawn items' identifiers/URLs are retained indefinitely.
- The metadata of withdrawn items will not be searchable.
- Changes to deposited items are not permitted.
- If necessary, an updated version may be deposited as well.

NOTE: In the event of CSIR Research Space being closed down, the database will be transferred to another appropriate archive.






































Attachment c: Structure of CSIR Research Space










































The structure of CSIR Research Space is as follows:






































- Community: Biosciences
 - Collections: Agroprocessing and chemical technology; Analytical science; Aptamer technology; Bioprospecting; Discovery chemistry; Enzyme technologies; Microbial expression systems; Plant biotechnology ; Structural biology; Systems biology; and Yeast expression systems
- Community: Built environment
 - Collections: Architectural sciences; Construction; Infrastructure engineering; Infrastructure systems and operations; Logistics and quantitative methods; Planning support systems; and Rural infrastructure and services
- Community: CSIR Publications
 - Collections: CSIR Annual Reports; CSIR e-News; and CSIR ScienceScope
- Community: Defence, peace, safety & security
 - Collections: Aeronautic systems; Landward sciences; Optronic sensor systems; Radar and electronic warfare systems; Safety and security; Systems modelling; and Technology for special operations
- Community: General research interest
 - Collection: General research interest
- Community: General science, engineering & technology
 - Collection: General science, engineering & technology
 - Sub-collection: General science, engineering & technology
- Community: Information & communication technology
 - Collections: Accessibility research; Earth observation technologies; Education, youth, gender; High performance computing; Human language technologies; Open source ; and Wireless technologies
- Community: Information Services
 - Collection: Information services

- Community: Laser technology
 - Collections: Laser materials processing; and, Laser physics and technology
- Community: Materials science & manufacturing
 - Collections: Energy and processes; Fibres and textiles; Manufacturing science and technology; Metal and metal processes; Polymers and bioceramics; and Sensor science and technology
- Community: Metrology
 - Collection: Metrology
- Community: Mobile intelligent autonomous systems
 - Collection: Mobile intelligent autonomous systems
- Community: Nanotechnology
 - Collection: Nanotechnology
- Community: Natural resources & the environment
 - Collections: Climate change; Coastal and marine systems; Ecosystems processes; Environmental and resource economics; Environmental management; Forestry and wood science; Mining and geoscience; Pollution and waste; Resource-based sustainable development; Sustainability science; Sustainable energy futures; and Water resources and human health
- Community: Space technology
 - Collections: Earth observation; and Satellite tracking, telemetry, command
- Community: Synthetic biology
 - Collection: Synthetic biology

ATTACHMENT D: COMPLETE LIST AND USAGE BY COUNTRY – nOV 2007

| | Domains/Countries | | Pages | Hits | Bandwidth |
|---|--------------------------|-----|-------|-------|-----------|
| ? | Unknown | ip | 51006 | 85754 | 4.82 GB |
|  | South Africa | za | 19622 | 33982 | 2.13 GB |
|  | Network | net | 5894 | 12270 | 942.23 MB |
|  | USA Government | gov | 4966 | 5094 | 96.37 MB |
|  | Commercial | com | 4803 | 9097 | 651.00 MB |
|  | Germany | de | 1721 | 2339 | 138.43 MB |
|  | Netherlands | nl | 1090 | 1570 | 89.85 MB |
|  | Australia | au | 973 | 2263 | 234.98 MB |
|  | United Kingdom | uk | 919 | 2292 | 294.93 MB |
|  | India | in | 804 | 2057 | 94.98 MB |
|  | Switzerland | ch | 751 | 896 | 52.03 MB |
|  | USA Educational | edu | 746 | 1859 | 99.68 MB |
|  | Poland | pl | 449 | 713 | 45.47 MB |
|  | France | fr | 444 | 883 | 45.74 MB |
|  | Canada | ca | 358 | 1003 | 95.49 MB |
|  | Japan | jp | 264 | 493 | 46.20 MB |
|  | Italy | it | 262 | 521 | 39.07 MB |
|  | Indonesia | id | 233 | 389 | 57.68 MB |
|  | Portugal | pt | 230 | 550 | 34.33 MB |
|  | Sweden | se | 218 | 458 | 20.87 MB |
|  | Turkey | tr | 199 | 314 | 28.49 MB |
|  | New Zealand | nz | 192 | 399 | 9.44 MB |
|  | Brazil | br | 190 | 430 | 44.40 MB |
|  | Non-Profit Organizations | org | 179 | 531 | 19.07 MB |
|  | Thailand | th | 163 | 357 | 28.52 MB |
|  | China | cn | 153 | 324 | 19.22 MB |
|  | Israel | il | 151 | 265 | 20.47 MB |
|  | Slovenia | si | 149 | 211 | 9.96 MB |
|  | Ireland | ie | 131 | 274 | 25.22 MB |
|  | Singapore | sg | 125 | 326 | 18.94 MB |
|  | Malaysia | my | 117 | 263 | 20.76 MB |
|  | Colombia | co | 113 | 260 | 8.68 MB |
|  | Belgium | be | 107 | 246 | 31.76 MB |
|  | Greece | gr | 104 | 252 | 16.98 MB |
|  | Spain | es | 102 | 256 | 48.05 MB |
|  | Pakistan | pk | 99 | 245 | 13.32 MB |
|  | Romania | ro | 96 | 145 | 29.03 MB |
|  | Argentina | ar | 95 | 187 | 11.21 MB |

| | | | | | |
|---|----------------------|------|----|-----|-----------|
|  | Botswana | bw | 89 | 124 | 15.62 MB |
|  | Finland | fi | 85 | 193 | 11.46 MB |
|  | Russian Federation | ru | 83 | 147 | 24.79 MB |
|  | Egypt | eg | 83 | 175 | 17.27 MB |
|  | Zimbabwe | zw | 81 | 155 | 26.62 MB |
|  | Taiwan | tw | 80 | 171 | 17.54 MB |
|  | Croatia | hr | 74 | 90 | 3.45 MB |
|  | United States | us | 67 | 193 | 15.46 MB |
|  | Ukraine | ua | 65 | 102 | 4.22 MB |
|  | Vietnam | vn | 64 | 92 | 13.40 MB |
|  | Old style Arpanet | arpa | 63 | 80 | 5.94 MB |
|  | Mexico | mx | 62 | 146 | 23.10 MB |
|  | Estonia | ee | 62 | 248 | 2.17 MB |
|  | Lebanon | lb | 60 | 68 | 2.46 MB |
|  | Bulgaria | bg | 60 | 60 | 7.30 MB |
|  | USA Military | mil | 60 | 132 | 25.94 MB |
|  | Czech Republic | cz | 52 | 80 | 6.24 MB |
|  | Hungary | hu | 51 | 114 | 10.49 MB |
|  | Lesotho | ls | 50 | 109 | 1.40 MB |
|  | Namibia | na | 48 | 78 | 10.53 MB |
|  | Norway | no | 48 | 111 | 20.07 MB |
|  | Lithuania | lt | 47 | 106 | 6.52 MB |
|  | Morocco | ma | 46 | 99 | 11.46 MB |
|  | Denmark | dk | 42 | 107 | 3.28 MB |
|  | Chile | cl | 36 | 99 | 11.88 MB |
|  | Unknown | adsl | 34 | 34 | 347.27 KB |
|  | Guatemala | gt | 32 | 54 | 8.42 MB |
|  | Trinidad and Tobago | tt | 32 | 60 | 6.43 MB |
|  | Tanzania | tz | 32 | 60 | 4.01 MB |
|  | Austria | at | 30 | 79 | 10.72 MB |
|  | Slovak Republic | sk | 26 | 68 | 2.80 MB |
|  | Swaziland | sz | 24 | 25 | 1.37 MB |
|  | Ghana | gh | 24 | 75 | 8.25 MB |
|  | United Arab Emirates | ae | 23 | 37 | 8.96 MB |
|  | Peru | pe | 21 | 56 | 9.08 MB |
|  | Saudi Arabia | sa | 21 | 35 | 10.79 MB |
|  | South Korea | kr | 18 | 55 | 1.42 MB |
|  | Philippines | ph | 16 | 37 | 3.23 MB |
|  | Uganda | ug | 15 | 33 | 1.52 MB |
|  | Mauritius | mu | 15 | 38 | 2.28 MB |
|  | Hong Kong | hk | 14 | 43 | 972.70 KB |

| | | | | | |
|---|-----------------------------|--------------|----|----|-----------|
|  | Oman | om | 14 | 35 | 405.63 KB |
|  | Mozambique | mz | 13 | 48 | 1.86 MB |
|  | Yugoslavia | yu | 13 | 20 | 3.99 MB |
|  | Uruguay | uy | 12 | 40 | 787.77 KB |
|  | Biz domains | biz | 11 | 25 | 1.20 MB |
|  | Rwanda | rw | 10 | 17 | 179.50 KB |
|  | Belarus | by | 10 | 10 | 2.91 MB |
|  | Samoa Islands | ws | 8 | 36 | 153.34 KB |
|  | Sri Lanka | lk | 8 | 29 | 398.15 KB |
|  | Ivory Coast (Cote D'Ivoire) | ci | 7 | 7 | 19.83 KB |
|  | Kenya | ke | 6 | 20 | 830.05 KB |
|  | Syria | sy | 5 | 5 | 678.24 KB |
|  | Cuba | cu | 5 | 5 | 3.28 MB |
|  | Luxembourg | lu | 5 | 5 | 508.80 KB |
|  | Zambia | zm | 5 | 19 | 790.32 KB |
|  | Bahamas | bs | 4 | 18 | 140.30 KB |
|  | Iran | ir | 4 | 4 | 1.41 MB |
|  | Moldova | md | 4 | 18 | 137.49 KB |
|  | Madagascar | mg | 4 | 18 | 137.37 KB |
|  | Malawi | mw | 3 | 3 | 299.88 KB |
|  | Bosnia-Herzegovina | ba | 3 | 3 | 1.58 MB |
|  | Latvia | lv | 3 | 3 | 6.50 MB |
|  | Jordan | jo | 3 | 10 | 93.89 KB |
|  | Virgin Islands (USA) | vi | 2 | 2 | 175.52 KB |
|  | Burkina Faso | bf | 2 | 2 | 25.48 KB |
|  | Nepal | np | 2 | 2 | 91.31 KB |
|  | Papua New Guinea | pg | 2 | 9 | 68.74 KB |
|  | Nigeria | ng | 2 | 9 | 70.77 KB |
|  | Venezuela | ve | 2 | 2 | 841.14 KB |
|  | Cocos (Keeling) Islands | cc | 2 | 9 | 70.16 KB |
|  | Eritrea | er | 2 | 2 | 79.92 KB |
|  | Fiji | fj | 2 | 9 | 69.79 KB |
|  | Bolivia | bo | 1 | 1 | 15.98 KB |
|  | Maldives | mv | 1 | 1 | 82.69 KB |
|  | Unknown | mtnnsn et | 1 | 1 | 137.69 KB |
|  | Kazakhstan | kz | 1 | 3 | 25.59 KB |
|  | Unknown | invalid | | 2 | 34.62 KB |