A DRIVER's Guide to European Repositories

A DRIVER's Guide to European Repositories

Edited by Kasja Weenink, Leo Waaijers and Karen van Godtsenhoven

Amsterdam University Press

This work contains descriptions of the DRIVER project findings, work and products. In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately via info@surf.nl.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the DRIVER project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this work hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the DRIVER consortium and the authors, and can in no way be taken to reflect the views of the European Union.

Publisher: Amsterdam University Press, Amsterdam
Cover design: Maedium, Utrecht
Lay-out: JAPES, Amsterdam

ISBN 978 90 5356 411 0
NUR 953

Contents

About the contributors

Authors

Dr. René van Horik works as a theme manager for DANS (Data Archiving and Networked Services www.knaw.dans.nl). DANS is the Dutch national organisation responsible for storing and providing permanent access to research data from the humanities and social sciences in the Netherlands. René is involved in research and projects to enhance the research data infrastructure in the humanities and social sciences.

Wilma Mossink, LL.M., is the legal advisor of SURFfoundation and SURFdiensten. Her expertise is in copyright management in higher education. She was project manager for two work packages in the SURF-JISC collaboration on copyright: Publishing agreements, institutional copyright policies and institutional repositories: bringing the Zwolle agenda to fruition. Wilma developed the Copyrighttoolbox which includes the Licence to Publish as one of its most important features. Another part of her work is the legal aspects of open access. She also drafts and comments on licences for use of content and software. Furthermore, Wilma advises the legal committee of the FOBID, the Dutch Library Forum. In this capacity, she represents the Netherlands in the Copyright Expert Group of Eblida, the European organisation for libraries. She is Dutch representative in the Copyright and Legal matters committee of IFLA (CLM) and is its information coordinator.

Vanessa Proudman has been a project manager for over 10 years in local, national and international information network/service projects for a UN-affiliated organisation and Tilburg Univ. library. She has been the Nereus Programme Manager since its foundation in 2003: a consortium of twenty academic libraries from leading institutions in the field of economics including Oxford, LSE, EUR, and Tilburg University. Since Sept. 2007 she has been project manager of the NEEO (Network of European Economists Online) EU-project. She is currently particularly interested in operational challenges surrounding repositories, service-development based on repository content, user-centred repository design, advocacy, and cost-effectiveness in these areas through knowledge-exchange initiatives.

Barbara Sierman, MA, studied Dutch literature and started her career in library automation at OCLC-PICA (Pica at that time) in 1979. After that she worked at several IT companies as a consultant. In 2005 she started her job as digital preservation officer at the Koninklijke Bibliotheek (National Library of the Netherlands). Barbara specialises in preservation plan-

ning and preservation metadata and has contributed to several (international) working groups and publications.

Dr. Alma Swan has a PhD in cell biology from Southampton University and an MBA from Warwick Business School. After a lectureship at Leicester University and a senior managing editor position at Pergamon Press (later Elsevier Science), she jointly founded the consultancy Key Perspectives Ltd. in 1996, which undertakes business development and market research work in the scholarly communications arena. Alma is a business strategy tutor for Warwick Business School's MBA programme, a business mentor/teacher for Southampton University's School of Management, Visiting Researcher in the School of Electronics & Computer Science at Southampton University, and Associate Fellow in the Marketing & Strategic Management Group at Warwick Business School. She is an elected member of the Governing Board of Euroscience (the European Association for the Promotion of Science and Technology).

Editors

Karen Van Godtsenhoven, MA studied English Studies, Comparative Literature and Library and Information Sciences. She started working on copyright projects for Ghent University Library in 2006, and then became the DRIVER project manager for Ghent. Within DRIVER, she is responsible for usability assessment, national networks and advocacy activities.

Dr. Leo Waaijers studied Mathematics and Theoretical Physics in Leiden. In 1964 an almost lifelong career followed at TU Delft where he started as a scientist, including a Ph.D. in mathematics in 1968. In 1977 he switched to management, at first as the personal manager of his department, to become member of the University Executive Board from 1984 to 1988. Following discontinuation of this position he was appointed University Librarian. In this function he realized the new library building, the merger with the university press and the transfer to the new library system Aleph. In 2001 he accepted a corresponding post at Wageningen University & Research Centre. As off January 1-th 2004 he is manager of the SURF Platform ICT and Research.

Kasja Weenink, MA studied History in Groningen and has worked as a researcher in Migration Studies at Amsterdam University. From 2003 until 2006 she has worked at the Ministry of Education, Culture and Science in the Netherlands, first as an Auditor, later as a Policy Advisor on Scholarly Information issues. She started working as a project coordinator for the SURF Platform ICT and Research in 2006. In the DRIVER project she is responsible for the editing and coordination of the Focused Studies.

About the DRIVER studies

DRIVER, or the *Digital Repositories Infrastructure Vision for European Research*, is a joint collaboration between ten European partners which aims to create a knowledge base for European research.[1] DRIVER is funded by the EU (FP6) and puts in place a test-bed of digital repositories across Europe, to assist with the development of a knowledge infrastructure for the European Research Area. The project builds upon existing institutional repositories and national networks, from countries including the Netherlands, Germany, France, Belgium and the UK.

DRIVER engages itself to collect only publications that are open access. This means that the end-user, when performing a search, only retrieves records that contain full text, or openly available research data. DRIVER also prepares for the future expansion and upgrade of the digital repository infrastructure across Europe and ensures the widest possible user involvement. In order to stimulate the development of state-of-the-art technology and to harmonise European practices in this respect, DRIVER has executed a set of strategic and coordinated studies on digital repositories and related topics.

The European Repository Landscape by Maurits van der Graaf (Pleiade, Netherlands) and Kwame van Eijndhoven (SURF, Netherlands) inventories the present type and level of OAI- compliant repository activities in the EU. The study shows that in 15 EU countries a sizeable part of the research universities has implemented a digital repository for research output: in seven of these countries it is estimated that more than half of the research universities have done so. Yet, the study also shows that 5 five EU countries seem to be in a starting phase, and some countries do not appear to have any repository. Next to the issue of basic implementation of the repositories, the number of full- text publications in the existing repositories can be further improved. Van der Graaf urges universities and decision makers to accelerate current developments since free access to knowledge and research outputs are important drivers for the knowledge society.

A Driver's Guide to Repositories, edited by Kasja Weenink, Leo Waaijers and Karen van Godtsenhoven (SURF, NL and University of Ghent, Belgium), aims to motivate and promote the further creation, development and networking of repositories. It contains comprehensive and current information on digital repository-related issues particularly relevant to repository managers, decision makers, funding agencies and infrastructure services as stakeholders. DRIVER has identified five specific, complex and long-term issues which are essential to either the establishment, development or sustainability of a digital repository; the business of digital repositories, stimuli for depositing materials into repositories, intellectual property rights, data curation, and long- term preservation. The success of a

repository is dependent on having addressed these five issues sufficiently. Good practices and lessons learned as part of this report will assist stakeholders in both the institutional repository day-to-day and long-term challenges, and can help them to avoid reinventing the wheel. The study focuses on inter- and transnational approaches which go beyond local interests.

The *Investigative Study of Standards for Digital Repositories and Related Services* by Muriel Foulonneau and Francis André (CNRS, France) reviews the current standards, protocols and applications in the domain of digital repositories. Special attention is being paid to the interoperability of repositories to enhance the exchange of data in repositories. The study is meant for institutional repository managers, service providers, repository software developers and generally, all players taking an active part in the creation of the digital repository infrastructure for e-research and e-learning. It aims to stimulate discussion about these topics and supports initiatives for the integration and (in some cases) development of new standards. The study also looks at the nearby future: which steps have to be taken now in order to comply with future demands?

The production of the studies is being coordinated by SURF, the collaborative organisation for higher education and research, aimed at breakthrough innovations in ICT in the Netherlands, in close association with Amsterdam University Press and the following DRIVER partners: CNRS (France), the University of Ghent (Belgium), ICM (Poland), the University of Gottingen (Germany), the University of Bielefeld (Germany), UKOLN (University of Bath, UK), and the University of Nottingham (UK). The editors would like to thank Amsterdam University Press for the pleasant cooperation.

More information about the DRIVER project and publications can be found at www.driver-community.eu

1. Introduction

We can expect, within a fairly short time frame, that each research-based institution in Europe will have a repository and that the research outputs from each institution will be collected in and disseminated through the repository. Within the scope of this publication, a digital repository is being defined as

1. Containing research results,
2. Institutional and/or thematic, and
3. OAI-PMH compliant.[1]

Institutional repositories contain scholarly publications (reports, working papers, preprints, post prints and published versions of articles and books) produced by universities or research institutions. Thematic repositories are usually organised around a specific discipline or research domain. All digital repositories, either institutional or thematic, comply with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which enables their contents to be widely shared. Digital repositories contribute to the open access movement by providing platforms for researchers to make research results freely available on the web. They contribute to improving the visibility of research results, typically scientific articles, and are as such an important part of the digital repositories infrastructure vision for European research (DRIVER).

This *DRIVER's Guide to Repositories* aims to motivate and promote the creation, development and networking of digital repositories. The guide does not provide strict directions on how to construct a repository, or network of repositories. It contains comprehensive and current information on digital repository-related issues in the research community and is particularly relevant to repository managers, decision makers, funding agencies and infrastructure services as stakeholders. This guide not only supports the institutions that already participate in the current EU-funded DRIVER network, it also reaches out to institutions that are about to get started with repositories or aim to further extend their current services or impact.[2]

DRIVER has identified five specific, complex and longer-term issues which are essential to either the establishment, development or sustainability of a digital repository:

- the business of digital repositories
- the population of repositories
- intellectual property rights
- data curation
- long-term preservation

The success of a repository depends on having addressed these five issues sufficiently. Good practices and lessons learned as part of this report will assist stakeholders in both their day-to-day and long-term challenges, and can help them avoid reinventing the wheel. These issues will be addressed in five chapters, which all focus on inter- and transnational approaches.

In the chapter on the business of repositories, Alma Swan aims to support a) those who are planning to set up a digital repository for their institution or other organisation; b) those who have already established one and who would like a new perspective on certain issues, and c) those who are in the early stages of thinking about a repository but have not yet taken the plunge. In a pragmatic way, Swan guides the decision-making process for establishing a digital repository. Swan lists five business operational models that seem applicable to repository-related developments. What is important for the organisation is that the product or service offered is sensible, manageable and able to be realised within the resources of the organisation. Even a straightforward institutional repository with no frills needs careful thought and a plan for its implementation and ongoing management.

Vanessa Proudman investigates the challenges in populating repositories in Europe, taking a selection of case studies which reflect the types of digital repository and service models in existence. Six cases have been analysed:
- a university institutional repository (University of Minho)
- a university school repository, run by a research department, and a campus wide institutional repository run by a library which closely liases with its school repository. (University of Southampton)
- a central archive repository which brings together national research results (HAL)
- an international research organisation institutional repository (CERN)
- a subject-specific service model built on institutional repository content (Connecting Africa)
- a service which increases institutional repository content quality (Cream of Science)

An in-depth description of the six case studies can be found on the DRIVER website www.driver-community.eu These descriptions go into great depth about policy issues, organisational choices, issues surrounding the establishment of the repository or service, population mechanisms, take-up, services, advocacy, legal issues and sustainability. The chapter on stimuli for depositing material into institutional repositories focuses on the comparison of the six cases on the aforementioned issues surrounding the population of repositories. The chapter ends with seventeen pointers for stimulating the population of repositories which have been distilled form the case studies. Lists of critical success factors and inhibiting factors in populating repositories can be found at the aforementioned website.

Difficulties with solving copyrights problems often hamper the filling of digital repositories and hinder the smooth management of the repository. The recurring questions about intellectual and commercial ownership of the works in the repository take a lot of time and can even hinder the creation of a fully accessible repository. To help different stakeholders to overcome these copyright issues, Wilma Mossink's chapter on intellectual property rights contains concrete examples and models. The chapter starts with an overview of copyright and other intellectual rights relevant for digital repositories. It also provides insight into what work on intellectual property rights is already being done in the EU. Furthermore the study provides models to continue working with and to develop in the EU digital repository context in order to arrive at sustainable development and operation at the local, national and international level. The starting point of this study is the central position of the author in the landscape of scholarly information. It examines the legal relationships an author has to enter into to make his/her work fully openly available. Appendix 2 contains an overview of the relevant European initiatives and good practices. This part contains contact information for people in the EU who are engaged in the legal aspects of digital repositories.

The research and publishing processes are becoming more interwoven. New developments in the fields of knowledge sharing and dissemination blend together tools, research data and publications. Repositories do not only contain the traditional publications, but also pre-prints and related datasets. As we can see in the chapter on data curation by René van Horik several curation activities are required to maintain and preserve the digital research data as well as to facilitate the future reuse of research data. Data curation is a relatively new term and used within the context of a wide number of objects. Van Horik first elaborates on the data curation concept. Features of the scientific digital objects are then described which are relevant to data curation. The third part covers data quality issues. This chapter ends by addressing data curation tools, concepts and procedures.

The domain of long-term preservation is closely related to data curation. Barbara Sierman describes the development of this domain. In the last decade, many articles have been written and conferences have been organised around the theme of digital preservation. Despite all these efforts, digital preservation has not reached its full potential. Digital repositories have collections of scientific treasures which are waiting to be found by contemporary and future generations. It is now vital to find ways to preserve this valuable material in the long term and to guarantee access to it in the future. However, a clear recipe with rules and guidelines on how to carry out digital preservation in a consistent manner is not yet documented. The digital repository community is still searching for the best way to handle digital material for the long term. Sierman describes the different developments in this relatively new area, offering the different stakeholders a broader perspective, and drawing attention to the different relevant issues at stake and possible solutions.

To support the further creation, development and networking of digital repositories, the appendices of this guide offer additional information on relevant developments in the field of business models and intellectual property rights. This journey will be supported by the DRIVER project's website, www.driver-community.eu, which can guide the reader through the European repository landscape.

2. The business of digital repositories

Alma Swan

2.1 Overview

It will be surprising if there are any tertiary-level research-based or teaching institutions in Europe that do not have a digital repository within a few years. Worldwide, repositories have been increasing at an average rate of about one per day over the last year or so and this can be expected to gather pace further. The reasons for having a repository are so compelling, the advantages so obvious, the payoff so potentially large, that no institution seriously intent upon its mission, and upon enhancing its profile and internal functioning, will want to disadvantage itself badly by not having one (or more).

Digital repositories can also be developed and maintained by a subject community (or entity acting on behalf of a subject community). These are more usually established by harvesting content from institutional repositories, but there are a few exceptions where subject community repositories attract content from the creators directly. Institutional and subject repositories have many purposes in common, but institutions find additional, institution-specific advantages in having a repository, too. Digital repositories have a number of functions or foci:

- to ***open up and offer*** the outputs of the institution or community to the world
- to ***impact on and influence*** developments by maximising the visibility of outputs and providing the greatest possible chance of enhanced impact as a result
- to ***showcase and sell*** the institution to interested constituencies – prospective staff, prospective students and other stakeholders
- to ***collect and curate*** digital outputs (or inputs, in the case of special collections)
- to ***manage and measure*** research and teaching activities
- to ***provide and promote*** a workspace for work-in-progress, and for collaborative or large-scale projects
- to ***facilitate and further*** the development and sharing of digital teaching materials and aids
- to ***support and sustain*** student endeavours, including providing access to theses and dissertations and providing a location for the development of e-portfolios

This chapter covers the business issues around digital repositories – their raisons d'être, putting forth a business case for repositories, the costs and

resources associated with them, and the things managers must think about and plan for in sustaining and developing them. Repositories can cost a lot to establish, or very little. They can succeed in gathering huge amounts of content, or end up with hardly any at all. They can become part of the working life of an institution or their users, or they can be largely ignored by the population they are set up to serve. They can raise the profile of an institution rather spectacularly, becoming a true asset in its mission, or they can contribute to its obscurity. Those responsible for instigating and running a repository have much work ahead in managing it so that it successfully achieves the expectations of which it is capable.

We should remember, amidst all the excitement about repositories, that they are quite a new phenomenon. Apart from the few in the vanguard, most repositories have been established within the last four years or so. Moreover, they are evolving rapidly as technologies develop and as the ways in which researchers and learners – and administrators – accommodate to the digital age and its opportunities. Much has been learned already about how best to develop successful repositories but we need to keep sight of the fact that things change and develop and improve all the time. What is considered good and useful today will be surpassed by something very good and more useful next year. It is an exciting and challenging working scene for those involved.

This chapter aims to set out describe those aspects of that scene that pertain to setting up and running a repository. It provides a formal framework for thinking about the purposes of repositories and how they can offer an improved scenario for many aspects of scholarly communication and assessment. It describes the types of business model – ways of running a repository – that are most appropriate to institutions within academia, and it discusses the issues that repository managers need to take into account in order to give their repository the best chance of success in the short and the medium term. Beyond that, none of us can look. We live in fast-moving times that are seeing not only massive technological developments but also the shifts in attitude and behaviour that characterise the 'netgen' – the generation that has grown up with the Internet and the world wide web. Indeed, one of the challenges for repositories would seem to be that their relative formality contrasts with the informal, more spontaneous and very attractive opportunities for communication offered by blogs and wikis. That is something to which we will need to pay attention as time goes on.

Repository services are one of the main keys to success for repositories, and this chapter also deals with their business models. Useful, popular services can really boost the use of repositories, both by information creators and information seekers. Repository managers need to ensure the content of their repository is fully visible and harvestable by service providers who will drive the use of that content as a result. They also need to ensure that there is some content there to be harvested.

A number of managers of established, successful repositories have been consulted for this study. Their experiences and opinions are reported to help readers gain from real-life cases. Their practically accumulated wis-

dom will be much more useful than my theory-based analysis, though there is some of that, too, where it seemed appropriate. The chapter reflects what we currently know about best practices in the business issues around establishing and running a repository and hopefully it will be a useful aid for those who wish to progress along that path.

2.2 Digital repository developments in Europe

There is much interest in developing and promoting digital repositories for research information in Europe. Strategically, a network of repositories offers the basis for the *Single Information Space* and the *European Research Infrastructure* objectives of the European Commission with the attendant promise of huge benefits to the research community of Europe and to the European population as a whole. Digital repositories collecting and housing the outputs of European research will provide the infrastructure for communication between scientists, for technology transfer between the research community and industry, and for the wider aim of improving the links between science and society as a whole. Repository developments, through improved accessibility and communications, are expected to lead to benefits in the environment, education, healthcare and economic well-being of the people of Europe.

At the time of writing, a study (e-SCI-DR) is underway that has been commissioned by the European Commission's Information Society and Media Directorate General. The study will identify the e-infrastructure required for e-science digital repositories and will provide the Commission with an overview of repository developments in Europe and set out the key issues. We can expect substantial advances in the field of digital repositories as a result.

On the ground, the DRIVER project that has spawned this volume is promoting the establishment of digital repositories by research organisations across the continent.[1] And preceding DRIVER, two national-level repository network developments were already in place. In the Netherlands, the DAREnet network encompasses a repository in every Dutch university.[2] In the UK, the SHERPA project supports and encourages the establishment of digital repositories in UK universities.[3] There is a brief overview of the business models of these repository networks in section 2.11. Similar developments can be seen in other countries.

The digital repository network will keep company in Europe with the pan-European GEANT network, funded under the Fifth Framework Programme and focusing on connectivity, and with the Grids infrastructure, funded largely under the Sixth Framework Programme and focusing on information processing. Together these form the integrated e-infrastructure that will enable new ways of working, most importantly that commonly referred to as 'e-science', the establishment of virtual collaborative research groups both within and across disciplines. The European Com-

mission has indicated in the past that it has as one of its goals the further integration of projects and developments in this area, with a scope which is pan-European and beyond the boundaries of existing project consortia or specific fields or disciplines.

These enabling mechanisms will be complemented by the distributed digital repository network being developed by research institutions and research communities, the focus of this book. We can expect, within a fairly short time frame, that each research-based institution in Europe will own a repository and that the research outputs from each institution will be collected in and disseminated from the repository. Research outputs comprise not only research publications, but also supporting data sets, conference contributions, working papers, theses and other item types, all available on an open access basis. The vision of the Single Information Space is on the way to becoming a reality.

There are a number of key issues around how repositories can successfully provide this basis for the advancement of research, scholarship, learning and technology transfer. Setting up a repository is only the start of the process and is relatively easy in the overall scheme of things. Once established, there are challenges in collecting content, in looking after that content in the face of the ever-changing digital information world, in adding value to the content and maximising its usefulness, and in ensuring that the bases on which repositories operate are legally sound. The other chapters in this book deal with these issues and provide timely and accurate information for repository managers and institutions. Here I deal specifically with the business issues involved in planning, setting up and operating a digital repository.

2.3 The context, and some definition

This chapter is aimed at people who are planning a digital repository for their institution or other organisation, those who have already established one and who would like a new perspective on certain issues, and those who are in the early stages of thinking about a repository but have not yet taken the plunge. There is much to learn from the experiences of those who are in the vanguard of repository developments and data and information collected from operating repositories are reported here to draw conclusions that help to take things forward generally and specifically. The other constituency that may find something of use here comprises the managers of actual or potential ***repository services***, entities that operate on repositories to enhance value and provide new offerings to users.

Since the term 'business model' can be applied in a variety of ways a clear definition of what this chapter is all about seems the optimal way to start. Before the web, businesses applied a functional model from a comparatively restricted range: they traded to maximise revenue; or they traded

to optimise revenue whilst pursuing professional goals; or they traded while pursuing a non-profit business mission. In all these cases things were rather simple and in all of them there was some sort of exchange of goods or services for money somewhere along the line.

With the advent of the web, e-business became a possibility for the first time and with it a whole raft of new ways of doing business emerged. As complexity has grown, so has the range of definitions of the term 'business model'. I don't want to dwell on this too much, or to turn it into an academic exercise, but in our context here there is some merit in finding a way to settle on a suitable method of scoping what I shall be dealing with in this chapter. Our context here is one where, unlike in most other business situations, revenue generation assumes a back seat. That is not to say it is not involved at all, nor that it may not become more central in the future; rather it is to say that, *currently*, revenue generation is not high on the list of priorities where digital repositories are concerned. And let us for the sake of clarity state here that we are talking about *research community digital repositories* and that our coverage does not extend to the digital collections created and managed by commercial or non-commercial publishers.

One of the most formulaic (and most useful in general business contexts) definitions of a business model is that put forward by Chesborough and Rosenbloom, who provided a list of six factors that a business model encompasses, as follows:[4]

- articulation of the value proposition
- identification of a target market segment(s)
- definition of the business's value chain
- specification of revenue-generation mechanisms
- specification of the business's position within the value network
- formulation of the business's competitive strategy

These are spot-on for any new trading business formulating its strategy for the future, but do they help us think about models for digital repositories? The answer is that some elements do, and I will discuss these later. Meanwhile, I suggest that for repository managers planning and framing the scope of their activities, the pragmatic approach of Clarke, discussing business models for open source software enterprises, is the most relevant as well as being the easiest to work with. He defined the issue as a series of questions:[5]

- who pays?
- pays what?
- for what?
- to whom?
- why?

This definition covers everything that is pertinent to business modelling for repositories, as the rest of this chapter tries to make clear. You may think there is still an overemphasis on money even in this business model definition, but if you are an existing or potential repository manager this issue

will undoubtedly be quite near the forefront of your concerns. And, as we shall see, it is central but doesn't have to be dominant.

The last thing to be said in this introductory piece is that a business model is very definitely not the same as a business plan. To implement a successful repository there has to be an additional question at the end of Clarke's list – How? That is where the business plan comes into effect.

2.4 The value chain

Businesses analyse where they sit in the value chain associated with their business activity. Elements of value are identified and analysed in relation to the offering in hand. For trading businesses, the value proposition is made to their customers. For scholarly digital repositories, the value proposition is made to the scholarly community.

Readers will be familiar with the concept of the ***scholarly communication value chain*** – the set of activities that enables content created at one end of the process to be delivered to its audience at the other. The actors in the chain are content creators (scholars), reviewers, publishers, intermediaries (e.g. subscription agents), libraries, navigation and discovery services, document delivery services, rights management services and so forth.[6] The scholarly communication process has been described as having four main elements:[7]

- registration: the establishment of priority on an intellectual creation (an idea, a concept or research finding)
- certification: the validation of the quality of the intellectual effort or of the research finding
- awareness: the ensuring of the accessibility, availability and dissemination of intellectual and research outputs for others to build upon, and
- archiving: the storage and preservation or intellectual or research outputs as an intellectual heritage for future users

For the present purpose I propose a somewhat longer list of elements that comprise the value chain. We can then use this to compare the value to the user offered by the traditional providers of that value – academic publishers – with that provided by digital repositories. The outcome is most clearly shown by a value curve and this is presented in figure 1. The four elements above are there, but I have split the 'awareness' one into its constituent parts and added others, so that the full list is:

Registration:	the establishment of priority on an intellectual creation (an idea, a concept or research finding)
Certification:	the validation of the quality of the intellectual effort or research finding, usually done by peer review
Availability/dissemination:	making research outputs available to users (which is different from accessibility)
Accessibility:	the ease with which users can get access to available outputs
Cost to user:	how much cash the user has to part with to gain access to available outputs
Navigability:	the facility for searching, finding and retrieving research outputs
Look and feel:	the quality of presentation and utility of outputs
Additional functionality:	extra value that is added, such as citation linking, adding context, linking to supporting data, etc.
Editorial value:	copy editing, translations, reproduction
Usage feedback:	data for the user (author) on how the output is being read, cited, used, incorporated into the progress of science
Preservation:	the storage and preservation or intellectual or research outputs as an intellectual heritage for future users

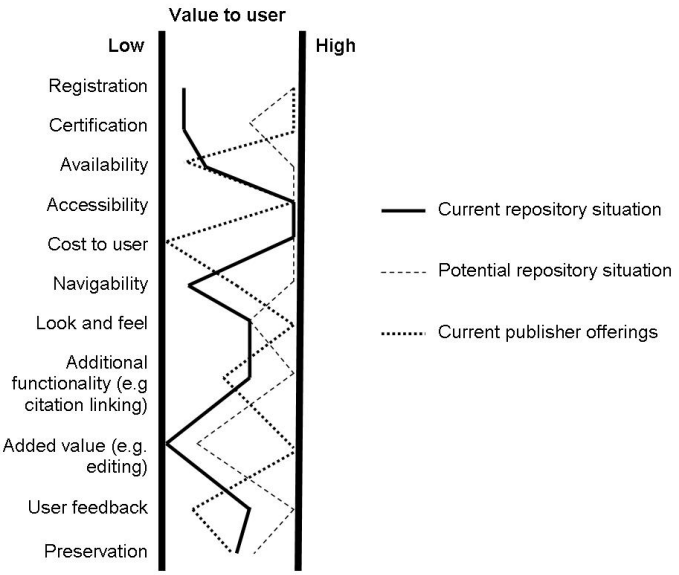

Figure 1 – The value curve for research communication

Figure 1 actually shows three value curves: one for digital repositories with their current offerings; one for digital repositories with potential offerings;

and one for the existing offerings from scholarly publishers. Such a value curve graphically demonstrates where the value lies (and how much of it there is) for each element of the value chain. It shows where digital repositories achieve maximal value for the user, where they don't and where they might do so in time.

The curve is largely self-explanatory. Repositories currently play little role in the ***registration*** process (except for the physics arXiv, which scientists do use as a location for announcing and laying claim to new findings), nor in ***certification*** which remains part of the formal publishing process (though there are moves in this direction).[8] With some exceptions, repositories also do not increase the ***availability*** of research outputs over and above their availability through conventional means. What repositories do well, though, is to maximise the ***accessibility*** of available outputs, at no ***cost to the user***, hence the high-value scores for both of these factors. At present, ***navigation*** tools for repository content are rather rudimentary compared with those provided on a commercial basis for the published literature, though this will change as more sophisticated services are developed on top of the repository network. The same applies to the ***functionality*** and ***presentation*** of repository content, both of which are basic at the moment but ripe for development. And at the moment, except in those few examples where publications are produced from repository content, ***editorial value*** added by repositories is very low. This is not to say that in time editorial work might not be carried out on repository content: this is entirely possible, either at the level of individual repositories or, even more likely, over the repository network as a whole. For this reason, this element is given a high *potential* value. ***Usage reporting*** is already a feature of many repositories and some far-reaching developments in this area mean that in the future detailed analysis of usage, including the provenance of downloads and citations, will be a function of repositories, individually and on a collective basis. Finally, preservation and curation activities carried out by most digital repositories are fairly simple and consist mainly of maintaining digital files in their deposited formats. More advanced ***preservation and curation*** skills are going to be needed over time and these are likely to be provided by specialised services operating on a high-level (probably national or supra-national) basis.

2.5 The value proposition from repositories

Having analysed the value chain, businesses then articulate a value proposition to the customer, built upon that analysis. There are a number of perspectives from which the need for a value proposition for digital repositories can be viewed. For institutional and other open access repositories, the value proposition can be summarised as follows:

On behalf of the research community, a digital repository proposes to:

- maximise the accessibility,
- maximise the availability,

– enable the discoverability,
– enable increased functionality,
– enable long-term storage and curation, and
– enable other potential benefits of scholarly research outputs at no cost to the user.

The value proposition above is the one that repositories make to the wider research community. They do so from a position of commitment to the knowledge commons and to sharing the outcomes of publicly funded work. A second value proposition has to be put to the institution by the persons responsible for instigating the idea of the repository. Usually, the library or the IT department is responsible for this and must put a convincing case to senior management for establishing and running a repository, something that will require considerable resourcing from the institution over time. This is discussed in section 2.8. There is a third perspective to this, too. Institutions are not the only stakeholders with an interest in digital repository possibilities and a commitment to sharing the outcomes of publicly funded work. Research funders, from small players such as specialised charities to those at the highest level (for example, the European Union), have a vested interest in seeing their funding turn into results of some sort – progress in disease management or cures, improved applications, increased knowledge transfer, better innovation. The value chain associated with these sorts of imperatives is different to that of the scholarly communication value chain in figure 1 in the sense that it has additional elements and contexts, but the value proposition that ensues shares many characteristics with that given for digital repositories above. In particular, the 'other potential benefits' would include such issues as enabling the transfer of knowledge between sectors in the 'knowledge triangle' (research, education and industry) and maximising the efficacy of technology transfer.

2.6 A typology of business models for repositories and related services

In the last decade a number of authors have attempted to develop a typology of business models for web-based businesses.[9] In a study on repository services we reduced the extensive lists produced by these authors to a simpler list of five operational models that seemed applicable to repository-related developments.[10] These are:

1. institutionally owned: institutions own and run the business to further their own goals and strategies;
2. public bodies sponsor the business for the public good;
3. the business runs on a community basis, sustained by the communities they serve;
4. the business runs on a subscription basis, selling products or services to customers paying cash.

5. the business runs on a commercial basis (other than subscription-based): a number of sub-types are covered by this term, for example an advertising model.

These models are equally applicable in this current context and Clarke's questions that frame his definition of a business model can be answered as shown in figure 2, which shows the typology of the business models in tabular form.

	Institutional model	**Public sponsors model**	**Community model**	**Subscription model**	**Commercial model**
Who pays?	Institution	Public body, e. g. ICT organisation or research organisation	Community members	Users	Users or advertisers
Pays what?	Cash	Cash	Cash and/or in-kind	Cash, at intervals	Cash at point of use
For what?	Staff, hardware, software, services	Staff, hardware, software, services	Staff, hardware, software, services	Service or product	Service or product
To whom?	Itself via internal accounting; suppliers if outsourcing any supply elements	Service/product provider	Service/product provider	Service/product provider	Service/ product provider
Why?	To further institutional aims	To further public good	To further community aims	To acquire the service or product	To acquire the service or product

Figure 2 – Typology for business models for digital repositories

Which looks most appropriate for digital repositories? Actually, all of them are appropriate and all are in use. The ***institutional model*** is the one most commonly used for institutional repositories, unsurprisingly, though the ***community model*** also applies in some cases where a number of institutions collaborate on a repository. An example of such collaboration is the White Rose consortium comprising the universities of Sheffield, Leeds and York in the UK. The ***public sponsor model*** is the one adopted in France, where the HAL (Hyper-Article en Ligne) repository platform is funded by the Centre for Direct Scientific Communication (*Centre pour la communication scientifique directe,* CCSD) of CNRS, the national science funder.[11] The ***subscription model*** – if I may be permitted to stretch the definition a little – is repre-

sented, for example, by repositories that lease space or hosting facilities to other institutions that pay annually for the service. Tilburg University and Southampton University's School of Electronics & Computer Science do this. The ***commercial model*** is exemplified by repositories that offer additional, one-off, paid-for services such as digitisation or the sale of electronic theses. The University of Utrecht, for instance, does the latter.

So whilst it is true that the majority of digital repositories are operating on a non-commercial basis so far, the way has been shown for revenue generation, at least in a limited way, by offering expertise and services that others are willing to pay for.

Repository managers, or those aiming to become such, may well be considering whether such models might be adopted by their own organisation. There is much scope for such offerings as the European network of repositories expands: not all institutions or organisations wishing to have a repository function will want to take on all the tasks associated with running such an entity and will be happy to outsource all or part of the enterprise to a third party (or parties).

What is important is that the product or service offering is sensible, manageable and within the resources of the organisation placing it before the community. Even a straightforward institutional repository with no frills needs careful thought and a plan for its implementation and ongoing management. The next section helps thinking in this direction by disaggregating the general repository business model into its constituent parts and assessing what is involved in each, using the real-life examples wherever possible.

2.7 Components of the business model

We have the series of questions provided by Clarke: who pays, for what, to whom, how much and why? For a would-be repository manager there needs to be a clear answer to each before settling on a business model and developing a business plan, but those questions are too big on their own. We need to break things down into manageable chunks. In figure 3 the overall picture – that covered by Clarke's five questions – is represented in a matrix that aids analysis. The factors along the top are those that contribute to the general long-term prospects for the business; those down the left side are the activity areas for the business. There is a question at each intersection to indicate what is involved.

	Viability	Sustainability	Adaptability
Business case	Does our business offering fit stakeholder needs and preferences? **A**	What are the likely costs? **D**	Is our model adaptable and flexible? **G**
Business scope and development	Can we develop and launch this? **B**	Do our resources at least match the likely costs? **E**	Can we build in resilience? **H**
Business management	Can we manage this business? **C**	What resources can we find? **F**	Will all stakeholders remain committed? **I**

Figure 3 – Business analysis matrix

The **viability** factors are focused on making the business happen; the **sustainability** factors are concerned with the resourcing implications of the business; the **adaptability** factors are about future-proofing the business.

So that this can be applied in a practical situation, we need to expand the contents of each cell, and try to answer the resulting questions. To make this exercise useful in a practical way the managers of eleven European digital repositories answered a series of detailed questions that I put to them about their repository operations. There are thus some real-life examples and data to draw upon. One of the repositories represented is at the national level, harvesting some of its content from smaller, institutionally based repositories. One is a repository in a large university department. The rest are institutional repositories in the strict sense of the term.

Cell A: Where business case meets viability: Does our business fit stakeholder needs and preferences?

- Will the service fit stakeholder needs?
- Can we make the case to the institution/organisation?
- Is a pilot project necessary or advisable? Will it tell us much?

Cell B: Where scoping the business meets viability: Can we develop and launch this?

- What is the business going to offer?
- How might this change over the short to medium term?
- Can we do it all ourselves?

Cell C: Where management of the business meets viability: Can we manage this business successfully?

- What key performance indicators should we use?
- What goals might be thrust upon us by others?
- Do we need to outsource anything?

– How are we going to market our business?
– What new tasks might be involved?

Cell D: Where business case meets sustainability: What are the likely costs?
– What cost schedules are we likely to face?
– How do these fit with our medium-term budgets?
– What other resources might be needed and can we supply them?

Cell E: Where business development meets sustainability: Do our resources at least match our likely costs?
– Can we afford this business?
– Where might costs change?
– How does the resource implication of the business fit with our medium-to-long term plan?
– Can the costs be predicted (and met) in the medium term?

Cell F: Where business management meets sustainability: What resources can we find?
– Does our long-term plan allow this expenditure?
– What margin for error should we factor in?
– Can the goalposts be moved (and by whom and for what reason)?
– What potential exists for a change of business model?

Cell G: Where business case meets adaptability: Is our model adaptable and flexible?
– Can we build in flexibility?
– At what cost?
– Can we measure payoff?
– What new demands or goals may arise?

Cell H: Where business development meets adaptability: Can we build in resilience?
– What can we foresee?
– How will we cope with that?
– How will we monitor for future movements that might be significant?

Cell I: Where business management meets adaptability: Will all stakeholders remain committed?
– What new stakeholders might be brought in?
– What is the potential for new developments of any kind?
– What new national or international developments may have an impact?

This approach is mapped onto the business analysis matrix diagram and is shown in figure 4.

	Viability	**Sustainability**	**Adaptability**
Business case	Does our business fit stakeholder needs and preferences? – Will the service fit user needs? – Can we make the case to the institution/organisation? – Is a pilot project necessary or advisable? – Will it tell us much?	What are the likely costs? – What cost schedules are we likely to face? – How do these fit with our medium-term budgets? – What other resources might be needed and can we supply them?	Is our model adaptable? – Can we build in flexibility? – At what cost? – Can we measure payoff? – What new demands or goals may arise?
	A	**D**	**G**
Business scope and development	Can we develop and launch this? – What is the business going to offer? – How might this change over the short-to-medium term? – Can we do it all ourselves? – Can we make the case to the institution and to the users?	Do our resources at least match our likely costs? – Can we afford this business? – Where might costs change? – How does the resource implication of the business fit with our medium- to long-term plan? – Can the costs be predicted (and met) in the medium term?	Can we build in resilience? – What can we foresee? – How will we cope with that? – How will we monitor for future movements that might be significant?
	B	**E**	**H**
Business management	Can we manage this business successfully? – What key performance indicators should we use? – What goals might be thrust upon us by others? – Do we need to outsource anything? – How are we going to market our business? – What new tasks might be involved? – What policies and procedures need to be in place?	Is our model adaptable and flexible? – Does our long term plan allow this expenditure? – What margin for error should we factor in? – Can the goalposts be moved (and by whom and for what reason)? – What potential exists for a change of business model?	Will all stakeholders remain committed? – What new stakeholders might be brought in? – What is the potential for new developments of any kind? – What new national or international developments may have an impact?
	C	**F**	**I**

Figure 4 – Business analysis matrix developed further

The following three sections tackle the issues highlighted in figure 4. The information from the repository managers surveyed is used in answering many of the questions so that the answers are rooted as far as possible in real experience. The issues are considered under the three main headings – viability, sustainability and adaptability.

2.8 Viability of the repository

Main issues:

Stakeholder needs and preferences
Can the business be developed and launched?
Can the business be successfully managed?

2.8.1 Stakeholder needs and preferences

User requirements and needs

Repository stakeholders come in a number of guises – institutional managers, research managers, research funders, repository managers, end users (as authors) and end users (as readers).

- ***Institutional and research managers*** have an interest in marketing the institution, in providing a showcase for its activities, and in having an effective research management tool.
- ***Research funders*** want to be able to track the outcomes of their investments in research programmes and projects.
- ***Repository managers*** want to create a repository that is fit for all these purposes and can be managed within the resource constraints imposed upon them.
- ***End users*** **(authors)** need a home institutional repository that makes depositing their research outputs as simple as possible, that gives the best possible visibility and exposure for their work to the outside world, that acknowledges and facilitates privacy where necessary, that provides a collaborative workspace and that provides them with timely and accurate data on how their material is being accessed, read (downloaded) and used (cited or acknowledged).
- ***End users*** **(readers)** need a system that gives good findability, navigability and retrievability for distributed repository content across borders and boundaries, and over time (preservation of content).

Pilot repository projects

Virtually all the repositories surveyed ran a pilot project before launching the repository proper. A pilot identifies what difficulties would be asso-

ciated with running the repository, enables testing of procedures and practices and helps to assess staffing needs. Two of the sites surveyed also used the pilot project to find out how to get content into the repository and to develop some advocacy models. Overall, pilots prepare the way for a smoother launch than would otherwise be possible. Some pilot repository projects were specifically informed by the findings from the TARDis project which set out to examine the critical factors for success in setting up a multidisciplinary institutional repository.[12] There are distinct differences in the needs of different subject communities and the means to address and manage these must be built into a repository business plan.

Many repository managers found it useful to make a formal assessment of this pilot project stage. Five of them did this with respect to workflow issues and four of them looked at content to recruitment and user attitudes. Others looked carefully at the staffing involved and the financial implications. Analysis of workflow enables repository managers to make modifications such as creating new buffer areas and new sort features to manage the throughput of information, and the creation of tools to monitor individual patterns of work for members of the quality assessment team. Monitoring the budget enables forward forecasting and this is especially important in situations where certain factors, for example research assessment procedures, may lead to an uneven demand for resources across the academic year. In addition, most repository managers seem to log user feedback and use that to help prioritise work for the future.

Making the business case

The case for a repository must be made to the institution or community that will own and sustain it. A number of business reasons may be behind the establishment of a repository. The main ones are listed below:

- increasing the visibility and dissemination of research outputs
- providing free access to research outputs
- the preservation and curation of research outputs
- the collection of research outputs
- research assessment and monitoring
- a place for teaching and learning materials
- the development of special (or legacy) digital collections

In almost all cases surveyed the library has been one of the entities making the case for a repository. In half the cases, impetus has also come from administrators or from academic departments and in a few cases from the research office. In one case the IT department championed the cause. No real difficulty was reported by most of the respondents, although in the case of the national repository bureaucracy and formalities complicated the process.

In justifying a repository it is critical to work out a case that best aligns the repository's business with the main priorities of the institution. For research-based institutions this means focusing on the benefits to the institution in having a tool that can increase the usage and impact of its re-

search effort, maximise the visibility of its outputs and provide a management information system for monitoring and assessing the research carried out in the institution. In countries that have a formal national research assessment scheme, institutional repositories will be a boon to collecting data and compiling returns and a case can be made based on this issue. A repository is also a space for collaborative working and a location for work-in-progress so for institutions where large-scale (e-research) projects are taking place it can be argued that a repository would provide the infrastructural support that such undertakings require. For teaching institutions, the advantages of a repository for teaching and learning purposes can be highlighted – a place for the creation and stewardship of teaching materials and for their access by learners. In such institutions, too, the need for a place to develop student e-portfolios is part of the argument.[13]

The argument for a repository is, of course, quite a new one. Institutions have become accustomed to fencing off substantial parts of their budgets for IT purposes over many years now, and digital libraries have usually been part of that thrust, but a repository that collects the digital *output* of an institution, rather than a service that collects digital *inputs* (electronic journals, books and so forth) is something of a novelty. In comparison to the whole IT budget, the cost of a repository will be very small: even in comparison to the institutional library's spending on digital inputs the repository costs will be minimal. There *is* a cost, though, and if repositories assume over time a position that is much more centre stage in the institution's life, which is what is expected, then a proper and realistic budget line needs to be created for them. In some institutions the library has been expected to create and sustain a repository out of existing resources – both in financial and staff-time terms – but this is not an appropriate expectation if the institution is serious about its mission.

A carefully prepared case to senior management will highlight the appropriate advantages of the repository to the institution (see section 1 for a list of suggestions), will detail expected expenditure over a number of years, and will emphasise that the payoff is not measured in financial terms; instead, payoff will be measured by:

- improved visibility of the institution
- improved impact of its outputs
- more effective 'marketing' of the institution
- better management of the institution's intellectual assets
- easier assessment of what the institution is producing and creating
- facilitation of workflow for researchers and teachers
- facilitation of collaborative research

A framework for helping to articulate the value of digital materials and the need to take active steps to manage them has recently been developed at the University of Glasgow in the espida project.[14] This tool may be useful for people thinking about a repository and needing to gain senior management commitment.

2.8.2 Planning and launching the business

What the repository will offer

Planning the repository is essential. In most cases surveyed the planning phase lasted up to six months, but in a few cases it took up to a year and, in two cases, longer than this. The implementation was a lengthier process in general, mostly taking a year or more, but a few repositories were set up in a shorter time, usually as a result of detailed forward planning. Planning involves not only the documentation of technical development work but of procedures and policies for the repository once it is up and running. This is discussed further in section 2.8.3

Decisions have to be made regarding what types of material the repository is going to accept. Will it accept and store all types of research outputs – journal articles, data sets, theses, books and book chapters, working papers, grey literature, works-in-progress, conference contributions and so on? What sort of file formats will be accepted and will all of these have a guarantee of preservation? Of the repositories surveyed for this study, most accept a very wide variety of item types and a good variety of file formats. Acceptance is not the same as guaranteeing to preserve them, however. Preservation implies some additional specialised work on repository content. Those interested in knowing more about this can find authoritative information from the PREMIS and PRESERV projects.[15] Preservation of unusual file formats or complex objects may be considered by most institutional repositories to be outside their remit. Such tasks may be seen as the responsibility of specialised repository services. Examples exist, such as the Arts & Humanities Data Service in the UK, the Royal Library in the Netherlands, the Royal Netherlands Academy of Arts & Sciences and so forth – specialised national-level services with highly developed expertise in preservation and curation of digital objects.

Short-to medium-term changes

Even though repositories may decline to take on this sort of specialist work, managers should plan for the likelihood of other new developments in the short to medium term, particularly in the form of stakeholder-oriented services. Evidence suggests that it is repository services that will determine the uptake and success of repositories within the research community and, of the repositories surveyed for this study, all have already implemented some services for their repository and have others in the pipeline. Most of them have some sort of search capability and also provide usage feedback. Two have implemented the means to use the repository for research assessment and two have enabled the publication of electronic journals from the repository. Of those not currently providing usage data most have this as a planned activity: research assessment and e-journal publication are planned in two other cases. Other services planned for these repositories are:

– RSS/Atom feeds
– metadata enhancement

– harvesting to create subject-specific collections
– easy export of publications to authors' home pages
– easy export of publications to author CVs, grant proposals, etc.
– interoperability with the institution's CRIS (Current Research Information System)
– establishment of a 'collaboratory' (collaborative research) infrastructure based on the repository

2.8.3 Managing the repository business

In-house or outsourced?

Given that repositories are likely to grow more complex in their content and structure and that repository services are a popular and a critical determinant of acceptance and adoption by researchers, the question is raised of whether repository managers will be able to create and manage all this in-house.

Outsourcing of elements of repository creation and management is one option. Of the repositories surveyed here, all host the repository themselves on-site. There is the option of having a third party host the repository off-site, though, and this option seems to be quite popular in general. It frees up management resources and obviates the need for cash investment in hardware and software.

A 'halfway house' – outsourcing the building of the repository but hosting it on-site – is also an option and one being taken up by a growing number of institutions. The advantages are that the institution does not need to provide the expertise required to create the repository, but only that required to manage it subsequently (quite a different skill set), and that the job is done quickly and expertly by professionals. And although a certain amount of cash outlay is necessary, this is a straightforward budget item, something that may not be so simple when a repository is built in-house. A discussion of actual costs for repository building and management is found in sections 2.9.1 and 2.9.2.

Performance indicators

The performance of repositories can be measured in various ways. There is not yet a set of norms, but repository managers are assessing progress in a number of areas. An appropriate framework may emerge in time, perhaps akin to the 5S (streams, structures, spaces, scenarios, societies) quality framework for digital libraries.[16] I would suggest that some suitable indicators are:

Content recruitment:

– percentage of annual current research outputs of different kinds (journal articles, conference papers, theses) deposited in the repository
– percentage of legacy outputs retrieved and deposited

– special collections digitised and stored

User awareness and involvement:
– measurably raised level of author awareness of open access
– measurably raised level of author awareness of copyright issues
– measurably raised level of author awareness of general scholarly communication issues and developments

Workflow practices:
– quality assurance procedures
– throughput times stable or improving
– forecasting procedures developed
– peaks and troughs anticipated and smoothed
– repository embedded in the institution

Financial discipline:
– annual budgets and financial plans drawn up
– monitoring process in place
– forecasting process in place

The repository managers surveyed indicated that the biggest challenges they have faced so far have been content recruitment and making faculty aware of and engaged with the repository. Communicating with researchers is not considered to be a difficult process but getting the issues across to them is. Other serious challenges have been dealing with copyright issues and integrating the repository with workflow and existing work structures. Copyright and content recruitment are dealt with in detail in other chapters in this volume.

The respondents consider that their greatest successes have been increasing the visibility of the organisation's outputs and in providing free and timely access to them. Providing long-term access to repository content is also considered to have been successfully achieved, though ongoing stewardship and preservation has posed more problems.

Repository policies

Repositories need policies. Those that operate without any formal endorsement from the organisation tend to flounder. A repository policy may cover a number of issues, the most important being what the organisation requires of authors and what the repository is going to house. More than half of the repositories surveyed have a formal written procedure stating what types of material and what file formats can be accepted and so forth. The same number have a written policy stating the institutional aims for the repository and what is expected of authors. In all cases this policy was reviewed and approved at senior management level within the organisation.

All evidence to date shows that without a firm policy in place on what authors are expected to do about depositing their outputs, repositories remain virtually empty; with a mandatory policy they are filled much more

effectively.[17] In recognition of this, institutions and funders are now beginning to develop mandatory policies that are designed to provide open access to at least some of the outputs from the research they fund. Five of the seven research councils in the United Kingdom now have open access policies, as does the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG), the Flemish Research Foundation (Fonds voor Wetenschappelijk Onderzoek-Vlaanderen, FWO), CNRS, INRA, INRIA, the European Research Council, and a growing number of other funders and institutions across Europe.[18]

Such developments are likely to continue to grow in number and should be welcomed. They will aid repository managers in recruiting content to the repository and in raising awareness of open access and repository issues within the research community. A recent study we carried out for the JISC in the UK showed that researchers who are familiar with open access and its benefits have often learned of it through their funding body.[19]

External influencers

In terms of goals that might emanate from outside the repository's organisation, the sorts of policies discussed above rank as one of the most important and most likely. At present, most (though there are early-day exceptions) such policies relate to research outputs in the form of journal articles and perhaps conference papers in the relevant disciplines. Only a few as yet relate to *research data*, but the signs are that policies on data will follow suit. Repositories will need to plan for this, and resolve any issues about accepting data in various formats and in potentially large amounts. Some disciplines generate large numbers of small data sets, while others generate enormous data sets that will present something of a challenge to repositories. It is clear from the experiences and developments from those in the vanguard of this issue that institutional repositories will need to work with third parties that have specific expertise in data archiving and curation if long-term challenges are to be successfully met.

Marketing the repository

Advocacy within the organisation is important if the repository is to be used as an internal resource. The advocacy efforts of repository managers focus both on recruiting content and on driving the use of the repository as a collection of research outputs. Both of these are dealt with in the chapter by Vanessa Proudman in this book.

Marketing the repository externally requires different measures. If the organisation wishes to promote the repository as a resource to the rest of the world's research community it must have all the usual web marketing tools in place – a good home page indexed by the web search engines, links to this from all the relevant pages of the institution's website, and if possible reciprocal links with other institutions. It should also be registered with the ROAR (Registry of Open Access Repositories) and OpenDOAR services, which maintain a worldwide listing of open access repositories and

provide statistics on the content of each repository and its growth, the software used and other related matters.[20]

The primary purpose of digital repositories, however, is to provide a seamless database of worldwide content, searchable by all. In this context, the best marketing tool available for a repository is to ensure that it is indexed by Google/Google Scholar and other similar web services. We know that over 70% of researchers use these services to look for work-related information and that the majority of referrals to a repository are from external search engines.[21] For driving usage of a repository, therefore, Google and its ilk cannot be bettered.

2.9 Sustainability of the repository

Main issues:

What are the likely costs?
Do the resources available match the likely costs?
Is the business model flexible?

2.9.1 Present costs

Set-up costs

The costs of setting up a repository have long been discussed and, on occasions, reported.[22] The cost can be from a few thousand euros upwards, depending on how ambitious the repository intends to be. To illustrate this the following tables, from a study we carried out three years ago, show the costs incurred by a range of repositories from those set up by average-sized research-based universities to the one established at MIT on project funding.

Institution	Set-up costs	Running costs
MIT, US (DSpace)	USD$1.8m grant 3 FTE staff USD$400,000 system equipment **Total USD$2.4-2.5m**	staff: USD$225,000 operating costs: USD$25,000 systems equipment: USD$35,000 **Annual running costs USD$285,000**
National University Of Ireland, Maynooth	grant to he computer science student for set up and customisation 6 months grant for €5,000 for server **Total €20,000**	1 FTE staff member for upkeep and maintenance **Total €30,000**
Queens Qspace CARL, Canada	software: free server space at institution programmer for 12 months: CAN$50,000 staff costs for advocacy work with faculty hardware: CAN$2,065 **Total CAN$52,065**	library staff: CAN$25,000 ITS staff: CAN$25,000 **Total CAN$50,000**
SHERPA: Nottingham, UK	software: free standard server: £1,500 installation 2-5 FTE days: £600 initial customisation 15 FTE days: £1,800 **Total £3,900**	maintenance absorbed within HEI costs: 5 FTE days per annum co-ordination and collection of material: £30,000 three-year update of hardware and software: 2-5 FTE days and £3,900 **Total £33,900**

Figure 5 – Comparative set-up and running costs of a sample of repositories[23] [note the currencies used are those of the examples]

The repositories cited here were all built by the institution. The repository managers consulted for this present study provided further figures. For an ***in-house built*** repository the average set-up cost for an institutional repository, covering hardware and software costs, was €9,250. Staff time in setting up a repository averaged 1.5 FTE.

Most repositories surveyed used open source software so this was free, and two institutions wrote their own software. For one of these, the effort remains uncosted in detail but staff time is estimated at 1 FTE for one year. For the other institution, which undertook very extensive software development work, the cost was estimated at €250,000.

Outsourced repositories hosted at the home institution cost around €7,000 to set up, and ***outsourced built-and-hosted*** repositories around €38,000.

The outlier in the study was the big national repository which is running on 12 servers and has provision for dozens of terabytes of storage. Develop-

ment took four software engineer-years and the repository also bought a licence for the software it uses (and has modified). Very few institutions will need this sort of provision in the foreseeable future. The table below shows the costs broken down in a little more detail for an example repository from the SHERPA project. The repository belongs to an average-sized UK research-based university.

Initial set-up costs €		**Technical support / maintenance €**		**Annual operating costs €**		**Article input costs €**	
software	0	HEI standard web service maintenance: three-year upgrade		staff salary	51,000	hours per week	17.7
Server	2,550	hardware	5,100			articles per hour	4
installation	1,020	labour	1,020				
customisation	3,060						
	6,630		6,120				7.58

Figure 6 – Set-up and running costs of an institutional repository, based on the experiences of the SHERPA project in the UK[24]

Running costs

The table in figure 6 gives some indication of the ongoing cost of operating the repository and inputting articles (a task which is done by repository staff, not authors, at the institution used as an example). The running costs of a repository are also highly variable depending upon the range of repository activities undertaken, but the average staff allocation in the surveyed group of repositories is 2.5 FTE.

2.9.2 Future costs

Repository managers need to plan for the possibility of increasing costs in the following areas:

- Software developments: about a quarter of the surveyed repositories have made minor modifications to the repository software already; half have made major changes, and two-thirds continue to modify the software on a frequent basis. In addition, all the major repository software suppliers will periodically upgrade their products, which will entail costs of some sort – either for in-house work or for consultancy services to effect the upgrade;
- Increasing content: funder and institutional policies will inevitably have an effect on content recruitment for repositories. Where repositories are well embedded in institutional workflow this may be absorbed without severe cost implications, but this applies only to a small number of repositories that have found a way to successfully involve researchers in the

deposit process and minimise the need for third-party mediators. Where mediation (usually involving library staff) is the norm for the deposit itself or for quality-control procedures subsequent to the deposit, then increases in content will necessitate the investment of more staff time. None of the repository managers surveyed expected staff numbers to decrease. Half of them predicted that staff levels would remain static for the foreseeable future and half expected them to rise. A study on the relevance to sustainability of a repository of mediated-deposit versus author-deposit has just been published;[25]

- Development of services for the repository;
- The position of the repository in the business cycle – repositories at start-up or growth phases are likely to encounter unseen costs, whereas maturing repositories can forecast their costs much more accurately. It should be noted that repository businesses as a whole are a new phenomenon and looking ahead ten years to forecast where they might be and what they will be doing is very difficult at this time.

2.9.3 Flexibility of the repository business model

Factoring in change

The last words in the previous section emphasised how problematic long-term planning is for repositories at the present time. It is difficult to alight upon a suitable margin of error and new demands are difficult to forecast. Certainly, planning should allow for growth and for continued advocacy and marketing for repositories. The other certainty is change, and so repository managers should be prepared for managing change in whatever manifestations it appears. The most likely areas where managers need to plan flexibility into their repository operations are:

- Deposit practice: currently, there appears to be a fairly even split between repositories that allow authors to deposit content into the repository and those where deposit is a mediated process carried out by repository staff. With increasing amounts of content a shift to author-deposit may be a pragmatic move, even if repository staff still need to carry out some level of subsequent quality control.
- Content types: the primary goal of most digital repositories in institutions is to collect and make accessible conventional research outputs. As we move towards the Single Information Space, however, repositories may be required to house many other content types, some of which may have special requirements.
- Metadata enhancement: this is an extremely active field with many developments occurring in it. Metadata will become more complex and refined and much of this will be executed by machines. Metadata is currently often enhanced by repository staff after authors enter basic-level metadata at deposit. Over half the repositories surveyed here have such a system in place. In addition, half of them import external metadata which are then mapped to the repository metadata format by system

programs (in a few cases this mapping is carried out by hand). This type of activity will grow and repository managers will need to take this into account in their planning.

Potential for changes in business model

There is also the potential for a change in business model as a repository matures. So far, most digital repositories have developed only a limited number of services and in most cases provide these free to users. A small number currently charge for hosting repositories for other institutions or organisations and a small number charge for providing access to masters or doctoral theses. There are many other services that could be developed, however. Some are potentially revenue generating, such as publishing services, current awareness services or services to particular research or teaching communities. One repository manager reported that the demand for services and the ideas for them from researchers in his institution had been overwhelming. This is an innovative and creative field and one that has huge potential at local level and on a global scale. It is important, in the context of maximising the number of effective services that can be developed, that repositories adhere to the DRIVER guidelines on OAI-PMH compliance when exposing metadata. The DRIVER documentation details the necessary work for implementation of this. Usable metadata mean that repository content can be successfully harvested by service providers; non-standard metadata can condemn an item to obscurity. Further illuminating discussion on this issue of 'marketing with metadata' can be found in one of the reports from the PerX project.[26]

2.10 Adaptability of the repository

Main issues:

Is the business adaptable?
Can resilience be built in?
Will all stakeholders remain committed?

A repository will need to adapt as technologies, user behaviours and external influences change, and all are likely to change considerably over the medium term. The pertinent issues are:

– Flexibility: a repository is flexible if it has the means to adjust to new norms and practices. One example could be research data: at present, very few data are routinely deposited by researchers but there is increasing interest and activity on this topic at policy level. Funders are beginning to discuss – and some implement – open data policies. Some journals, such as *Nature*, already have such a policy. As these developments

continue, repositories will be receiving greater volumes of data and, moreover, data of many different types and in many different digital formats. A flexible repository will be one where forward planning has taken such developments into account and where procedures and facilities are in place to cope with what might be quite a sudden shift in this area. For most institutions there will be the need to work in partnership with expert providers of preservation services for data that is not run-of-the-mill.[27]

- Resilience: a repository will be resilient if plans are in place for adjusting repository capacity, workflow, and – if mediated deposit is the norm – staffing. Additional demands, such as a change in the form of a mandatory policy from the institution, will bring new challenges for repository staff in increased advocacy and awareness-raising activities.
- Monitoring for future developments: 'horizon-scanning' capabilities are essential in a world where repository and scholarly communication developments are happening very quickly.
- New stakeholders: might new stakeholders appear? One example of such a thing is the research assessment procedures that are increasingly being put in place around the world. In the UK, for example, the periodic national Research Assessment Exercise (RAE) requires a resource-intensive process of individual expert review of research outputs from each researcher. In future, it has been announced, the RAE will be 'metrics-based', and development of the new metrics that will enable this is about to begin. Usage statistics will almost certainly be one of the metrics incorporated into such assessment procedures and institutional repositories are the most natural locus for measuring such usage. The body that runs the UK's RAE will therefore become an interested stakeholder in the UK's institutional repository network and its developments.
- Development potential: repositories will not stand still. The potential for developing services is great and experience shows that as users begin to use services they ask for more.
- Performance measurement: finding the right indicators to measure performance is going to be crucial. Currently, repository managers measure the number of items, and the number of full-text items, in the repository; they measure downloads, and they measure interest (internal and external). Additional, perhaps more granular, measures can be developed that assess how the repository is being used; crucially, the degree of embedding of the repository in the general life of the institution and its workflow patterns needs to be assessed. Sustainability comes with embedding, and embedding means the full adoption of the repository as an everyday workplace tool by researchers across the institution and, in this context, repositories must also become an everyday tool for research administrators.

2.11 Organised repository networks

Two national-level repository organisations – DAREnet[28] and SHERPA[29] – have already been mentioned in section 2.1. The business model that they have adopted is simple. At the ***data level*** are the repositories, set up and managed by organisations; at the ***services level*** value is added to develop services. The model for the DAREnet network, developed by the SURF organisation in the Netherlands, is shown in figure 7.

At the data level, institutions collect, store and retain control over their own intellectual property in digital form. This is important for the institutions and ensures that the provision research content remains the responsibility of the data provider sector. It is in the interest of institutions to provide such access to its own outputs and the national network simply organises and enables institutions to do this. At the service level, services may be developed at an institution for that institution, or may serve a national audience or even a global one. Some services may aid the ingest process: for example, there may be services that advise on intellectual property issues, on metadata creation and enhancement, on technology, on preservation, or offer repository hosting facilities. SHERPA DP (digital preservation) is an example of such a service; SHERPA's RoMEO and JULIET services provide information on publisher permissions and research funder open access policies respectively, helping repository managers and authors understand their rights and obligations with respect to making their work open access.[30] Other services operate above the data layer and offer things as subject portals, theses collections, or the collected outputs from a particular set of institutions.[31] Services may be provided through a variety of business models, as discussed in section 2.7.

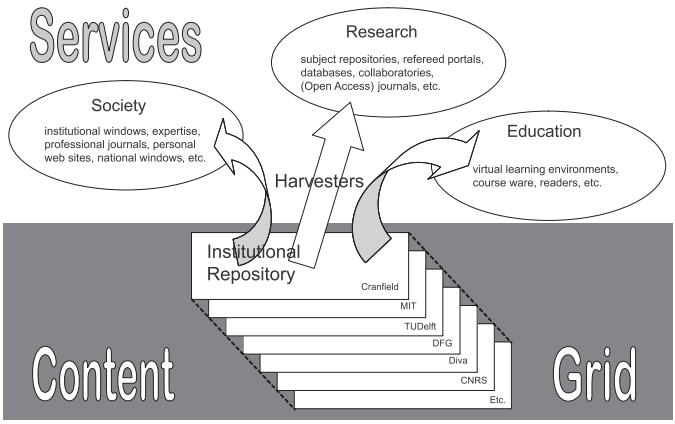

Figure 7 – The Darenet model for service and content provision

SHERPA has progressed on successive rounds of public funding through JISC in the UK. DAREnet has been funded over some years by the SURF organisation in the Netherlands.

2.12 Repository services and their business models

2.12.1 Business models

This section looks at the services that can be built around repositories. There is huge scope for developing repository services, adding value to the primary material collecting in digital repositories across the world. Repository services may adopt any one of a number of business models discussed in section 2.6, namely:

- institutional model
- public sponsors model
- community model
- subscription model
- commercial model

If there is a revenue-generating imperative one of the last two will be most appropriate. If not, or if the service is to be run for purely local (institutional) use, one of the other models may be most appropriate. It may be helpful to discuss this in terms of examples. Here I draw on previous work we did for a JISC-funded project on linking repositories and how repository services might fit into a national repository network scenario in the UK (see footnote 11).

Many repository services begin life as projects and the shift to sustainable service can take various forms. Some existing services have adopted a fully commercial model while others are in transition. In others only a publicly funded or community model looks workable. The possible general models for repository services are:

- **institutional model:** institutions own and run the service for institutional reasons, that is, that the service furthers their own goals and strategies
- **publicly funded model:** services that do not have the basis for revenue generation and are not appropriate for institutional or community models; these will need to be sponsored by public funding in the long term
- **community model:** services sustained by the communities in which they operate in a collaborative effort
- **subscription model:** services that can sell a product or service on a subscription basis to paying customers in the marketplace
- **commercial model:** services that can generate revenue in the marketplace

2.12.2 Types of repository service

There are already many examples of repository service in existence. Those related to ***ingest activities*** are:

- digitisation services: digitising legacy material such as older journal articles and theses, and special collection material
- IPR/copyright advisory/information services: advising on rights issues for authors, readers and institutions/repositories
- Open access advisory/information services: advising on issues around opening up research outputs of all types
- technical advisory services: advising on OAI compliance and similar technical issues around networking and interoperability
- repository building services: constructing repositories for organisations that wish to outsource this element
- repository hosting services: hosting repositories for organisations that wish to outsource this element

Services related to ***data-provision activity*** are:

- metadata creation services
- metadata enhancement services

The greatest opportunities for abundant and diverse services are where these relate to ***user needs*** and examples of these are:

- discipline- or subject-specific portals or current awareness services (e.g. ARNEX – Agricultural repository News Exchange)[32]
- access and authentication services: systems that integrate repository content with institutional records and databases
- usage data services: providing feedback on repository usage (downloads, citations, etc)
- preservation services: providing the expertise for long-term storage and curation of digital data
- research monitoring and analysis services: tools that enable the analysis of research outputs from an institution, set of institutions or larger
- resource discovery services: tools that enable the searching and retrieving of digital items within or across repositories
- personalisation services: gathering information of specific interest to specific users
- meta-analysis services: services that carry out national-level (or greater) analyses of research outcomes (e.g. for research funders)
- overlay journals: electronic journals developed from repository content (e.g. the Lund Virtual Medical Journal)[33]
- publishing services: peer review, copy-editing services and publishing services
- bridging services: services that map or point to repositories or their content for other services to use (e.g. the Information Environment Service Registry)[34]

2.12.3 Matching services and business models

In the table in figure 8 these service types are mapped onto the business model schema to show the models under which each service type might operate. The columns in the table are:

- Cost level:
  - ***Low (L)***: services with low running costs typically require low staffing levels and low levels of investment in fixed assets. For our purposes here, these are services that cost up to €170K per annum.
  - ***Medium (M)***: those that might cost up to €400K per annum
  - ***High (H)***: those with running costs above €400K per annum
- ***Appropriate business model(s)***: because there are multiple ways of making a business work some services have more than one appropriate business model.
- ***Scalability***: a score of 1 indicates that a service is highly and easily scalable, simply by incremental adding of the resources required. Scores of 2-4 indicate the need for some careful strategic business planning to scale up from a simple service to one satisfying more complex needs. A score of 5 indicates service that would be very difficult to scale up under its present operating model.
- ***Associated risks***: scalability is one thing that impacts on this but business risks arise from multiple sources such as change in the operating environment, in technologies and in the customer base and its requirements. Most ingest-level services are low risk. Those at output level that sell proven technologies come into this category. Medium-risk services may face scalability challenges but also face the challenge of continuing to match their offerings to a changing user needs base. As has been said earlier in this chapter, it is not easy to see far ahead. Digital information will continue to grow and technologies and their applications will continue to move very fast.

Service	Cost level	Appropriate business model					Scalability 1 = easy 5 = difficult	Associated risks	Comments
		Institutional	Publicly funded	Community	Subscription	Commercial			
INGEST SERVICES LAYER									
Digitisation services	M	✓				Merchant	1	Low	Institutions do their own digitisation, or pay a third party operating on a commercial basis
Rights/IPR advisory services	L		✓				1: but probably not required to scale substantially	Low	Core service
Open access advisory services	L		✓				1: but probably not required to scale substantially	Low	Core service
Technical advisory services	L		✓			Merchant	1: but probably not required to scale substantially	Low	Core service Some commercial operators may offer some as part of commercial repository-building service
Repository construction services	M	✓				Merchant	1	Low	Institutions do their own construction, or pay a third party operating on a commercial basis
Hosting services	M	✓				Merchant	1	Low	Institutions pay commercial operator

Service	Cost level	Appropriate business model: Institutional	Publicly funded	Community	Subscription	Commercial	Scalability 1 = easy 5 = difficult	Associated risks	Comments
DATA LAYER PROVISION									
Metadata creation and enhancement	M/H	✓	✓			Merchant Advertising	By machine: 2 By humans: 5	Medium	Existing and future JISC-funded projects may require long-term support Commercial companies will also operate in this niche
OUTPUT SERVICES LAYER									
Access and authentication services	H	✓			✓	Merchant	3	Medium	
Usage statistics	M	✓		✓		Merchant	2	Low	
Preservation	H	✓			✓	Merchant	5	High	Challenges will increase; Various models will operate for different user environments
Research monitoring	L	✓		✓		Merchant	2	Low	
Resource discovery	M/H		✓		✓	Advertising	4	Medium	Challenges will increase; Various models will operate for different user environments
Overlay journals	L	✓				Merchant Advertising	1	Low	Institutions can operate here (e.g. Lund Virtual Medical Journal), otherwise, lots of scope for commercial operators
Publishing	M	✓		✓		Merchant Advertising	1	Low-medium	Publishing services (e.g. peer review) may be provided on a commercial (publishers) or community (learned societies) basis. Value-added products may be produced on both bases too
Meta-analysis	L			✓		Merchant	2	Low	Development costs can be high but ongoing service costs should be low
Bridging services	M	✓	✓			Subscription Advertising	3	Medium	Core services Commercial companies may innovate in this niche

Figure 8 – Business models for repository services

3. The population of repositories

Vanessa Proudman

3.1 Introduction

This study investigates the challenges in populating repositories in Europe based on six good practices. These case studies have been selected to represent the current types of repository and repository service models in existence, with which most repository or related service managers can identify. Good practices have been chosen as a means to inspire the challenged or even disheartened. Cases have been analysed on a number of aspects such as policy issues, organisational choices, the establishment of the repository or service, population mechanisms, take-up, services, advocacy and legal issues. This contribution is intended as a guide to models which stimulate the population of repositories. The in-depth analysis of the cases on the DRIVER website www.driver-community.eu is of advantage to the reader who is seeking detailed information about a particular context.

This chapter has been primarily written for the repository manager. This can be an institutional repository manager, a departmental one, or a broader-reaching national or international disciplinary one; it can also be the manager of a service which has been established to bring together research from a number of sources. However, all managers have one thing in common: populating their archives is a challenge. A choice of solutions to the common problem will be provided and common guidelines to improve on population efforts conclude lessons learnt from these cases. The library director or information manager responsible for research information will also be interested in this study. It is important for these senior managers and policy makers to comprehend the complexities surrounding the challenges in populating repositories in order to make the necessary cultural changes. Higher-level European policy makers can use the results of this study for policy and funding programme development.

This study particularly focuses on the policy and organisational issues which have influenced deposit rates, which includes highlighting repository services. It likewise investigates take-up mechanisms on levels of senior management as well as within the research community. Various advocacy initiatives will be outlined which have supported these. Legal issues which either prevent or even stimulate the deployment of content are included in the analysis. These areas and their critical success factors and inhibiting factors for the population of the digital repositories and services are used as the basis for identifying concrete guidelines for better guaranteeing researcher take-up and content deployment in the future. It is hoped

that the in-depth investigation of these six cases will help others to tackle the issue of institutional repository population in their own environments, focusing on it anew within the broader context of other initiatives.

3.2 Method

3.2.1 Analytical framework

This study aims to investigate the stimuli for populating repositories. Based on the analysis of the case studies, six areas can be identified which influence the form and life of a populated repository: 1) policy issues, 2) organisation, 3) mechanisms and influential factors for populating repositories, 4) services, 5) advocacy and communication and 6) legal issues. These areas are also recurring themes of international and national discourse on the issue of open access and scholarly communication. They manifest themselves in various communication channels from appearing in influential documents such as SPARC position papers or monographs, as subject matter at open access and related scholarly communication conferences and workshops or as threads from related discussion lists or blogs.[1] However, what makes this study unique is that it goes into more depth into operational issues for the institutional repository manager, that is, the whys and hows, the critical success factors, the choices made and detailed contexts in order to be able to make informed decisions.

Six areas of analysis

It is **policy** which shapes the aim and the corpus of the repository's content. The Berlin Declaration on Open Access,[2] for example, mobilised a large number of leading institutions and their libraries to think differently about their role in scholarly communication. Harnad and Sale are strong advocators of the world of the 'green road to publishing'.[3] The ideal green road is a world in which each institution has the responsibility to implement broad institutional mandates to deposit academic output or else a place where patchwork mandates on individual or faculty levels are established.[4] However, discourse in countries such as France and the Netherlands for example suggests that other methods are necessary to encourage the participation of researchers where institutional mandates could be more difficult to enforce. This is where incentives need to play a more significant role.

This study therefore analyses institutions with mandates and looks at their successes. It will also investigate service models without mandates to deposit, but which aim to find their own niche to answer some of the problems of the researcher as author and reader. Operational aspects such as how a policy is implemented, how it is formalised and what is specified will also be addressed as a guide to those who have the intention to implement a policy or further develop it.

Organisational aspects that have an effect on the population of a repository will also be addressed. What types of driving forces can be utilised to implement policy? How important is high-level support in establishing an institutional repository and how can that stimulate the population of an archive? SPARC's position paper also recognises the challenges in organising academic output on an institutional level reflecting the embodiment of the research process and output; this will also come to light in this study.[5]

Further operational aspects of interest will be addressed by analysing **the mechanisms and influential factors for populating repositories**. Morag Mackie describes strategies for populating repositories in her article 'Filling institutional repositories: Practical strategies from the DAEDALUS project'. This paper looks at efforts involving library support as well as more sustainable ones by integrating deposit into the workflow of the researchers at the University of Glasgow.[6] In a similar manner, the six case studies for this research have been analysed by looking at the strategies which have been used to obtain material giving a small keyhole view onto some of the practices employed. This study will analyse a number of key areas which are crucial to better comprehend the acquisition of material for a repository or service. Researcher take-up, workflows, methods of content deployment, ingestion, content choices, i.e. current versus retro-digitisation, content type, and versions. For example, as Rowlands and Nichols point out in their international survey of scholarly communication of senior researchers, no one-size-fits-all solution is suitable for all disciplines. This chapter will look at the significance of considering disciplines when forming and evaluating our repositories.[7]

However, it is still a fact that the researcher is reluctant to post material to an institutional repository where the benefits are unclear. Davis and Connolly report on this whilst sharing experiences at Cornell.[8] Gierveld, in her article 'Considering a Marketing and Communications Approach for an Institutional Repository', sees the institutional repository as a product to attract a market of information providers. This means that authors and their needs need to be made central in order to attract the content necessary for populating an institutional repository.[9] The answer to this is a strong **advocacy** programme. This chapter will therefore list some of the concrete advocacy decisions made by the cases analysed, the role of advocacy in the repository's life, and the various methods and products in place and experimented with. After all, even despite an institutional mandate of her own in place, Callan claims that promoting, prodding and providing support to the author and potential institutional repository contributor is also essential for the successful population of a repository.[10]

In answer to addressing the target information provider more effectively, user-centred **services** are increasingly under development. SPARC's position paper talks of the content and service layer of the repository where value-added services not only assist in registration and certification, but also raise awareness of the repository's contents and facilitate its use.[11] In his position paper 'From libraries to 'libratories', Waaijers emphasises the great opportunity for libraries in developing services as part of their reposi-

tory package taking on a new role in scholarly communication.[12] This chapter will investigate the breadth of services made available by the cases in question by categorising them in areas of benefit to the researcher such as increasing visibility, information discovery and retrieval, and preserving access to research. The repository as a research assessment tool is also a potential service of significance to the researcher where repositories can serve as a CRIS (Current Research Information System). Day, in his report on institutional repositories and research assessment, claims that institutional repositories can potentially support research assessment in helping to generate, provide and store information on research output. However, he did not believe that e-Prints UK could deliver at the time of going to press in 2004.[13] Things have changed since, and this chapter will demonstrate that.

However, it is the **legal aspects** of self-archiving and open access which are one of the greatest challenges for populating repositories. SPARC's position paper and JISC's *Disciplinary Differences Report* point out that a clear lack of awareness of copyright and self-archiving is evident.[14] SPARC stresses that continued education on these issues will be necessary to better secure the regular deposit of quality content into repositories. One of the priority issues of the European University Association and its working group statement on open access is to promote the strengthening of the author's legal right to non-exclusive copyright and promotes the utilisation of model copyright agreements.[15] This chapter will similarly explore whether such IPR concerns are valid for all the cases studied and what influence this is having on the population of repositories and their services, and what mechanisms are in place to try to overcome these challenges.

3.2.2 Selection of the case studies

This study addresses six European case studies which demonstrate where the population of digital repositories in Europe is gaining ground. The first milestone in the research was to determine the case studies for study. Desk research was carried out using the directories OpenDOAR and ROARMAP to analyse the size of repositories as according to metadata and full text numbers as well as to observe growth patterns and rates.[17] As a result, a preliminary shortlist of European repositories and services was created. Initial telephone interviews were then carried out with those on the preliminary shortlist to verify the ROAR and OpenDOAR data on policies and metadata and full text statistics. Further questions were posed on growth and take-up. This resulted in the final selection of six case studies. All cases are also OAI-PMH repositories for inclusion in the DRIVER portal and address scholarly output. Cases are neither data archives, nor learning object ones. This study profiles repositories (institutional or broader) and services which stimulate digital repository population.

Success indicators which were used to determine case study selection were full text numbers, percentage of academic output, striking and/or steady growth data, and take-up by the research community. Central to the

selection of the case studies was differentiation along population policies, organisational profiles, repository types and services, language content, and geographical distribution. On the basis of these criteria, six cases were identified as good practices. The examples identified have been chosen to represent different models of repositories and services which have shown to stimulate institutional repository population. They are an inspiration for others and are to be seen as examples, and results of research carried out in 2006. They are neither typical nor completely unique. The following repository and service models have been identified:

1. A university institutional repository (University of Minho). Minho has been chosen due to the broad take-up from its research community, its interesting advocacy and support infrastructure and, above all, to highlight the implementation of mandates and incentives and their effects on repository content.
2. A university school repository, run by a research department (ECS Southampton) and a campus-wide institutional repository (Southampton) run by a library which closely liaises with its school institional repository. Southampton is known for its innovation in the area of scholarly communication. The relationship between a faculty-run repository and a newer campus one is of interest. The ECS archive belongs to the first OAI repositories.
3. A central archive repository which brings together national research results (HAL). It is profiled to analyse this different organisational model's approach and its results. It has no specific disciplinary focus as do many other known central repositories.
4. An international research organisation institutional repository (CERN) with authors from a tradition in self-archiving.
5. A subject-specific service model built on institutional repository content (Connecting Africa). This is a disciplinary service which is fed by a number of repositories. It serves a specific international community of researchers by providing a portal pulling information together to enhance networking and research.
6. A service which increases the quality of institutional repositories (Cream of Science). This is an example of a service which is built on a number of institutional repositories. It pulls together leading national researchers and their output in a decentralised manner, thereby showcasing national research results.

Due to the scope of the DRIVER project, this study is limited to six cases in the European domain. Good practices and lessons learnt in the repository field should know no geographical boundaries; it is clear that there are various interesting developments in Australia, Latin America and the United States. Due to the scope of this study these could not be addressed here.

3.2.3 Interviews

Once cases were selected, interviews were held face-to-face with library directors, institutional repository managers and initiators as well as with staff on more operational levels in some cases.[16] Interviews were held in late 2006 and early 2007, and therefore the data are based on 2005 and 2006 figures. In-depth case studies were then written based on a number of topics of relevance to the key target groups of the study, institutional repository managers and their policy makers. These detailed write-ups are available at www.driver.community.eu. By using good practices, cases are compared with one another, where relevant visible similar contexts or areas of interest and allowing one to see the variety of solutions to the population challenges there are at hand.

3.3 Good practices

This section provides executive summaries of the six investigated cases.

3.3.1 Minho University Institutional Repository[18]

Minho University is an example of an institutional repository within a university of 1,100 researchers and teachers, 13,500 undergraduate students and 2,000 graduate students. Research and education is carried out in the institutes of arts and humanities, social sciences, child studies, schools of sciences, health sciences, nursing and law, as well as some areas of engineering. At Minho University, the library director, the university's rector and the vice-dean of research all share a conviction of the necessity to make Minho's research materials open access, and it is their backing which has been crucial in establishing the institutional repository. Since 2005, Minho has consequently had a mandate in place requiring deposition of academic output, with an emphasis on currency and peer-reviewed material, coupled with a financial reward system to stimulate deposit until 2006 (100,000 euros) with a further 30,000 provided in 2006. These policies have brought about a dramatic increase in deposit rates, seeing a total of 2,813 documents deposited in the first year of implementation. The true effects of such financial incentives will come to light once deposit rates have been monitored after their termination. The library director's ambitions for the institutional repository are also partly owed to an investment in comprehensive knowledge exchange and advocacy activities on local, national and international levels.

The library has targeted all areas of Minho research. Its strategy has been to focus on departments where research decisions are made rather than at a faculty level, promoting their engagement by involving them in the decision-making process and giving them the autonomy to decide on the type of material they wish to showcase. These strategies, combined with policy,

seem to have been successful considering the 87% take-up from Minho's research departments. However, although departments have set up communities showing institutional commitment, researcher take-up is less pronounced in some areas. Some researchers from the department of bio-engineering for example contribute 80 to 90% of their findings to the institutional repository, whereas others do not. Minho institutional repository content deployment is managed decentrally in the form of departmental input via DSpace communities with local coordinators responsible for providing basic deposit support. The library provides a stable and reliable support structure to departments by supplying a broad range of information, training, help tools, and guidance in intellectual property rights (IPR) issues. This has also contributed to departmental take-up.

Minho is aiming for 100% coverage of its current publications. At present it is achieving 40% coverage of its annual academic output. However, Minho's library envisages that its potential new important role as the university CRIS, with the further implementation of tools to assist the researcher in the research process and the further analysis of incentive-building mechanisms, will reap even more results. This is still a challenge, although current policies have contributed to seeing the institutional repository being embraced by the majority of research management teams.

3.3.2 University of Southampton Research Repository, and School of Electronics and Computer Science (ECS) ECS EPrints Repository[19]

The ECS departmental institutional repository is an example where a university research and teaching department takes on its own self-archiving and institutional repository. This school contains 131 academics, 243 postgraduates and 213 research staff. The relationship between a leading departmental institutional repository and a later campus one (Soton) made up of 1,500 active research staff with its different challenges, modi operandi and missions, is also of research interest in this study. Soton is an example of a university institutional repository with challenges in integrating information available in various forms from across its campus. It serves a faculty of engineering, science and mathematics, a faculty of law, arts and social sciences, and a faculty of medicine, health and life sciences. The Soton institutional repository is highlighted here as an example of a repository with an important evaluation function serving as the main tool for the UK's Research Assessment Exercise (RAE).

The ECS institutional repository was established in 2001 and has been depositing under a mandate for several years. The school manages its own content and sees it as the responsibility of its authors to comply where advocacy and support is therefore limited. It is reaching 67% full text coverage of its present academic output. EPrints was developed and is maintained at ECS which can be seen as an asset in the implementation or adaptation of new services.

The Soton institutional repository, however, is run by the library. During the JISC-funded project TARDIS, which had as its scope the campus-wide implementation of an institutional repository, the library liaised with ECS and the EPrints team with the view to further focus on the needs of a broad range of future self-depositing researchers across the campus.[20] As a result, the Soton repository deposit system was simplified and tailored to researchers' needs and further services were developed in order to better sustain and manage future content deposition. The library currently invests considerably in quality control activities to ensure the access to quality content unlike its compatriot at ECS. Soton's library is responsible for aggregating information into its institutional repository from across the campus for the RAE, and although the mandate to deposit research results has been in place since 2005, it has seen a threefold rise in full text deposit since the year prior to its establishment with a population of 5,529 full-text files recorded in 2005. This has consequently both had a positive effect on its position in the university and resulted in more content deployment.

3.3.3 CERN document server[21]

CERN is an example of an international research organisation which serves a specific subject community, in this case, particle physics. These scientists come from a tradition of self-archiving which was stimulated by arXiv.org in 1991, yet CERN is facing challenges in populating its institutional repository with clients who prefer to deposit with data or service providers elsewhere.

CERN has had an institutional mandate in place since its establishment in the 1950s, and a mandate to deposit electronic copies of material since 2003. The library is also responsible for storing management information and documentation for CERN. As a result, approximately half of its content is delivered via departmental input by support staff, its authors and researchers using CERN facilities, amounting to approximately 1,000 documents per year. The remaining half is not deposited at CERN, but with other archives or information services such as arXiv.org, which offers added value research benefits to the researcher upon deposit.[22] CERN has faced this challenge by reclaiming some of that material by harvesting content, including metadata and, in some cases, full text by arrangement, from approximately 90 external data and service providers. As a consequence, CERN now obtains an approximate total of 80% of its annual academic output despite little investment in advocacy. CERN is now targeting the remaining 20% through advocacy and tool development. The institution has a drive to secure the acquisition of open access material in the future. The CERN management and library unanimously agree to strive to make all of its future publications open access. This is reflected in a concentration in open access publishing activities rather than self-archiving ones.

3.3.4 Hyper Article on Line (HAL inter-institutional repository)[23]

HAL is an example of a repository model which promotes the central deposit of quality research output in a complex national research environment. HAL is an archive, based on the organisational model of arXiv.org, which is interdisciplinary in nature and focuses on storing, improving and maintaining access to the full text quality research output of France. It provides a broad scope of content which is not confined to any one institutional or network boundary. The central archive forms a pragmatic solution to a research area where authors and research centres have multiple affiliations and do not necessarily identify with any one institution. HAL is a tool for information dissemination, discovery, retrieval and archiving of quality open access research output. It seeks to federate efforts via its archive, with a view to cost efficiency in administration, author support and preservation

The researcher is its prime focus. However, HAL does now specifically target many of France's research institutions, be they universities, research institutes or academies, recognising that organisational commitment and researcher support is crucial to further boost content deployment to obtain the critical mass it seeks. As a result, in July 2006 an agreement was signed between four leading research organisations in France: CNRS, INSERM, INRA and 86 of the council of rectors of France's universities to work on realising and contributing to the HAL platform for French research.

HAL provides a range of services which serve researcher individuals, research groups, and institutions in the research process and in the management of their information. HAL presently (January 2007) sees 1,300 deposits per month. It currently holds close to 40,000 full-text documents. The HAL management hopes that its new formal alliances with some of France's leading research institutions will further contribute to the enrichment of its archive.

3.3.5 Cream of Science[24]

Cream of Science is an example of a service which can stimulate the acquisition of high-quality content for institutional repositories. This service model also has the potential to increase prestigious support for open access and institutional repositories from the research community which may ultimately stimulate further author participation. The Cream of Science service is a national showcase of leading Dutch research and its researchers. The concept was developed to improve on the quality of Dutch institutional repository content. It was also the notion of using champions, that is, leading Dutch researchers selected for Cream, which was used to encourage institutional repository take-up by researchers in the future. Fifteen organisations, including all thirteen Dutch universities, collaborated in reaching a common milestone to showcase the work of over 200 leading researchers from all disciplines online. A Cream search service as well as automated

publication lists exist which include references to the entire oeuvre of leading researchers' work and links to the full texts of open access material.

At the beginning of the project there were concerns about copyright issues standing in the way of aggregating and disseminating material for the project via open access, and that the targeted authors would miss the opportunity. Yet, reality proved otherwise: researchers embraced the idea, resulting in more interest from authors to participate than could be met as part of the project. Participating authors actively collaborated in part, with some intent on providing 100% of their work online, others delivering boxes of material for digitisation. This resulted in populating Dutch institutional repositories with a total of 27,500 full-text files within less than a year, 80% of which were journal articles. The interpretation of Dutch copyright law permitting the digitisation, storage and dissemination of pre-1998 journal articles without the consent of the publisher also meant that 20,000 documents were digitised and put online as a result. This was brought to the Dutch repository community by Wilma Mossink, although it was Leo Waaijers who originally initiated the idea.

The ambitious project saw challenges in interoperability, quality control and retro-digitisation in particular. However, it was valuable experience gained for all participants prior to the ensuing embedding period of the institutional repositories into their institutions. Cream was an innovative stimulus for bringing high-quality content into local institutional repositories. Consequently, other countries are seeking to implement the model elsewhere. However, it still has to be investigated as to how far Cream has been a stimulus to encourage more continuous repository deployment.

Cream of Science was a nationally funded SURF project with participants from DAREnet , the name of the national repository network of large Dutch academic institutions. The service is now updated and maintained by the Royal Netherlands Academy of Arts and Sciences (KNAW)[25].

3.3.6 Connecting Africa (CA)[26]

Connecting Africa is an example of a subject service model with an international profile which has stimulated institutional repository deposit in the Netherlands and abroad. It is a portal which contains global information on African studies and has been developed as part of two SURF projects entitled 'DARC' and 'DARC2'. CA is run by the Africa Studies Centre (ASC) and its library. This is an organisation which undertakes social science research and promotes the dissemination of knowledge and understanding of African societies. Such an organisation, with a broad network and well recognised in its field, is in a suitable position to run an international portal and expert showcase on African studies. The initial DARC project addressed the ASC's local problem of the lack of structured research management information on its academic output. It therefore analysed related workflow issues from the outset. However, with its networking capabilities and national and international profile the ASC then brought Dutch

and European African studies together to further serve its international research community and promote networking by creating a gateway to African studies. CA also includes automated publication lists and expert profiles.

Connecting Africa's portal provides open access to publications and data of importance to African studies. CA has encouraged African studies scholars to deposit in institutional repositories. Evidence has shown that in the Netherlands, where the aim was to obtain content from 25% of all Dutch African studies scholars, the target was doubled by achieving 52%. This brought new content to a number of Dutch institutional repositories through researcher deposits and retro digitisation. CA harvests its content from decentralised locations and selects relevant content using a self-developed 'post-harvest analyser' that uses predefined African studies-specific algorithms. In January 2007, it was harvesting 44 institutional repositories in total from the Netherlands and Europe, extending this to Africa in the future. CA consequently holds over 10,000 object files (i.e. text and image files). The service has not only had an effect on deposit rates but also on policy. The ASC's policy on depositing research results became mandatory in 2007.

Connecting Africa also made a point of involving African studies researchers in its design. These researchers seem enthusiastic about the principle of the service and support the idea of gaining further worldwide visibility, and collaborate by providing content to institutional repositories partly for that purpose, although it is unclear as to how far they use the service for their own research. It is necessary, however, to monitor the steadiness of content deployment over time to see whether the service becomes rooted in the research community and achieves its aim in providing more open access content.

Connecting Africa aims to continue building a digital library for African studies scholars accessible to all online. CA is currently further extending its content stock to include documents from policy makers, NGOs, and journalists, building bridges between research and policy. Although CA is now maintained as part of ASC library activities, realising new ambitions will more than likely need to rely on project funding once again, which brings challenges in meeting the needs of its end users.

3.4 Learning from six European good practices

3.4.1 Policy issues

Policy issues are the backbone of repositories and their content stocks. Not only does policy establish repositories, but policy contributes to developing archive missions. The implementation of mandates and incentives to deposit material are critical factors here. These developments can bring about cultural change and can significantly contribute to gaining more population

results. Other funder mandates can hope to have an influence on population stocks. Content policies in particular will flavour the scope of the repository, its content type and numbers which need consideration before evaluating the success of a repository. Networking and knowledge exchange are also issues of importance for further development of repositories and their services. Policy brings focus and definition to repository efforts.

Policy development

Based on the cases studied, institutional policy proposals by and large come from those responsible for the repository archives. In one case studied however, Minho, it was the rector himself who proposed policies to use financial incentives to better guarantee current campus-wide content deployment. Such policies on mandates to deposit or other major strategic issues such as preservation, for example, have been presented to and endorsed by high-level management. At Southampton, for example, the University Research Policy Committee is the sounding board and decision-making body, although the ECS repository was endorsed by the school it serves. CERN reports to its Scientific Policy Board that also advises the library on policy development and operational issues. The rector is involved with the development of the Minho institutional repository and the senate of the university was involved when implementing the Minho mandate. For services such as HAL and Cream of Science, policy was established and developed by representatives of the institutional partners contributing their material. HAL established a specially assigned committee for strategic development in 2006 for policy development in the future, further representing contributing partners.

Mandates to deposit

Mandates to deposit electronic copies of academic output have been established to achieve the ambitious 100% academic output aims mentioned in most of the cases in this study. Such mandates have often been coupled with incentives, be they financial or ones that assist working practices. Mandates have undoubtedly had a positive effect on population statistics and several cases from this study, namely CERN, Minho and Southampton, demonstrate this. The establishment of a mandate to deposit by an institution is a clear signal that an institutional repository is a priority for institutional management. Minho's library director claims that the main factor for the successful population of an institutional repository is the establishment of an institutional mandate to deposit academic output. Minho's policy stipulates that researchers must and should deposit all research material and make it open access where possible. This was enforced in 2005. It was combined with a financial reward system which gave points to research departments dependent on the currency and version of the material deposited. The deployment of recent material and post-prints saw higher points gained. Its institutional repository's population consequently increased by 800% the following year, reaching almost 3,000 documents. Accompanying financial incentives seen at Minho can help obtain significant content

within a relatively short space of time for critical mass, causing a snowball effect. In 2006, financial mandates came to an end and deposits will need to be monitored over a period of time to obtain a fair picture of the mandate's influence on deposit rates. However, if the evidence shown by Sale is correct, deposits will continue to rise regardless.[27] One goal of HAL's operational head, D. Charnay, is to see a mandate to deposit in HAL, following Minho's success. However, this would not mean an institutional mandate but a mandate to deposit in an inter-institutional centralised archive. This could have far-reaching consequences for the French research community, impacting their organisations and the storage of their knowledge.

Southampton's mandates to deposit have seen increase in deposits since their implementation. This is true of both the mandate to deposit at ECS, established in late 2002, and the Soton campus institutional repository's 2005 mandate to deposit which supports the Research Assessment Exercise (RAE). ECS saw full-text deposits of 7% of its annual academic output in 2001 prior to the mandate, compared with 67% in 2006. Soton, on the other hand, had 311 full-text deposits in 2004 prior to its mandate to deposit; this increased to 1,208 deposits in 2006. As a result of the national agreement to only store digital object identifiers (DOIs) in institutional repository metadata for the RAE, which threatens the deposit of full-text content by researchers, Southampton proposed a new mandate to the university in 2006 which was implemented in 2007.[28] This institution-wide mandate stipulates that all journal article post-prints or publisher PDFs, where allowed by publishers, be stored in the repository to better ensure open access to its full text.

However, CERN sees the importance of a mandate as relative. Despite having a policy of depositing copies of research in place since the early 50's, as well as a mandate to deposit electronic copies of academic output since 2001, its community of researchers deposits 50% of its output into the CERN institutional repository. Other important research groups have shown to prefer to deposit elsewhere with established services, such as arXiv.org, which have proven to serve their research needs over time. CERN's repository manager, Jens Vigen, sees a mandate as a supportive tool to try to persuade those who do not deposit.

The involvement in e-information services can have a positive effect on policy development. The ASC director – initiator of CA – now wants complete coverage of his institution's output in the institutional repository and CA partly as a consequence of this service. He has established a policy to mandate the deposit of academic output by ASC's researchers in 2007 and is now considering sanctions for those who do not submit. Services can therefore motivate institutions to push for the deposit of material into their own repositories for example, to ultimately be presented in services of international and academic significance.

Mandates to deposit certain output with a well-defined authorship are increasing in the area of theses. Minho has already structurally been collecting electronic copies of all of its theses and dissertations since 2005. Soton is also seeking to mandate the deposit of electronic theses. The Netherlands

has seen some mandates to deposit electronic theses, and these in turn find their way into Dutch repository-based services such as Connecting Africa for example. However, institutional mandates in the cultural setting of the Netherlands are difficult to implement. This is where service development and research incentives have been, and still are, of the essence.

In summary, mandates can significantly contribute to the population of digital repositories as can be seen at Soton, CERN and Minho. However, evidence has also shown in these cases that mandates cannot be relied upon alone to achieve the 100% goals. Most people interviewed for the cases, whether or not they had already formed an opinion on mandates or services, in fact stipulated that both mandates and incentives are needed to achieve content deployment goals.

Incentives to deposit

Cream of Science and Connecting Africa have shown that services can be developed which can have a driving effect in populating repositories where mandates have a small part to play at present. The Cream of Science service has partly contributed to the recent 100,000 population target of all thirteen Dutch universities and two leading research centres. The Netherlands generally had no mandates to deposit in place at that time, apart from a few requiring the deposit of theses. Cream of Science was developed as an idea to attract leading researchers to deposit in repositories and to enhance the quality of repository content, as well as a means to showcase Dutch researchers and their research. At the outset libraries feared there would be little take-up from the research community, however, more researchers were interested in participating in Cream than capacities would allow. As of January 2007, 18 months after the end of the project, Cream still has a waiting list of scholars wanting to join the project. As a result, Cream of Science saw the deposit of 27,500 full texts, 80% of which were articles, during the nine months of the project. However, these numbers cannot compare with numbers of other repositories due to the large retro-digitisation of approximately 20,000 full-text documents. This service has significantly contributed to obtaining a critical mass of quality content, which has in turn seen more current academic output deposit in the Netherlands.

Connecting Africa focuses on a specific scientific community of African studies, with researchers dispersed across various institutions and their departments. CA started out as connecting African studies scholars in the Netherlands. However, it soon extended its goals to put that research in the context of more worldwide research; it now harvests European repositories, and African institutional repository harvests are planned for the future. Connecting Africa provides access to 10,799 object files, of which over 1,000 are articles, including many images of ethnographic and anthropological significance to the African studies community. Connecting Africa is an example of a disciplinary-based service which has seen increased author up-take and repository numbers rise as a result of the service. Just as with Cream, CA's numbers, should not be compared with those of other repositories mentioned in this study as it also contains retro-digitised material in

addition to current content, and does not focus on achieving annual academic output figures.

Services such as CA and Cream should be seen as stimuli for increasing the population of repositories. Material can be contributed in the realms of a particular initiative or project, which can then have further repercussions in the deposit behaviour of the researcher in the future.

Repository services need to be developed to answer a researcher's real research interests or problems such as increasing research impact, visibility and access to material. For both of these services, increased visibility has played a role: for Cream, researchers were profiled as leading lights in Dutch research, and for Connecting Africa, researchers were similarly profiled as experts with further international visibility of their work via a new African studies portal. These services are not limited by institutional boundaries, but are rather about extending networks and knowledge exchange in various disciplinary fields spanning various geographic levels. National showcases of leading research can certainly contribute to the quality population of repositories, but it is the disciplinary archives such as Connecting Africa or Economists Online (built on several institutional repositories) which will bring meaning to the contents of repositories.[29] This may well encourage further content deployment fulfilling local aims and at the same time the broader aim to influence and increase the impact of European research. These types of services can win the hearts and minds of researchers as authors and readers are more willing to contribute their work to repository efforts which are of scientific significance to them. This then supports information professionals in populating the institutional repositories. Smaller services which are built on repository content have similarly been developed for institutional repositories to enhance the deployment of further content, be there a mandate in place or not. For more information on these, see section 3.4.4.

Networking and knowledge exchange

All case interviewees stressed how vital networking and knowledge exchange was to their work. Knowledge exchange has proved to be important for policy and service development for the ultimate aim of populating repositories in several cases studied. Minho, for its own policy development, has looked abroad to the Massachusetts Institute of Technology and the Queensland University of Technology for inspiration.[30] CERN has a clear policy to continue working on international levels to develop added value services and archives for its particle physics community and to move forward on the road to 'golden' open access publishing, that is, the 'author pays' model.[31] However, this requires a commitment by those active in the open access community to share their experiences at international scholarly communication events – which all cases have done. Similarly, certain repository managers interviewed seek to play an active role in sharing their own activities to stimulate the population of their own repositories as well as to mobilise other institutions to do so.

Networking is clearly important for the successful population of services which are built on the collaboration with a multitude of authors and their organisations, as is the case with Connecting Africa, Cream of Science and HAL. For this reason, network products have been built into their services which are of potential interest to the network nodes. For example, Connecting Africa has done this with its expert pool and HAL allows the interrogation of its search system by searching for authors and/or disciplines to uncover networks of researchers and institutional collaborations.

SURF, as the main Dutch national funding body of institutional repository projects, considers professional networks as critical success factors for populating repositories. SURF set up a network of national repository communities consisting of targeted groups of decision makers, operational managers, technical and communication drivers who also drove Cream. Staff in similar functions engaged in the same types of issues and worked collaboratively on developing services and infrastructures for the improvement of Dutch repository population. SURF sees it as priority to continue with this model. At the time of going to press, other countries such as Germany, Portugal and the UK are considering following suit.

National funding bodies

Of the cases surveyed in Switzerland, the Netherlands, Portugal, France and the UK, in mid-2007 it was only the UK, which has seen a number of funding bodies, for example, the UK Research Councils, mandating the deposit of research output since 2006.[32] Others such as the NWO in the Netherlands have plans for implementation of such a mandate.[33] Such funding body mandates should further support repository population by promoting the discipline of depositing open access material into repositories. The Economic and Social Research Council (ESRC) of the UK, for example, has figures to show that full-text deposits have tripled from approximately 400 to 1,360 as of 31 March 2007 due to the ESRC's policy and mandate to deposit full text.

Policies on content

Content policies clearly influence repository metadata and full text numbers, which are the result of the mission of the repository or service. For this reason, population statistics need careful evaluation. For all repositories and the services studied, academic output is the targeted content. However, the type of content aggregated is dependent on the discipline the repository serves. For this reason, Minho and Soton have allowed departments to decide on the type of content to be delivered. At the ECS institutional repository, content is concentrated in the areas of journal articles and conference papers, which are most relevant to the scientific community of computer scientists. Academic output of an institutional repository, therefore, needs to be defined according to the academic scope of the institution in question.

All cases have a policy which strongly promotes the storage of full text. It should be pointed out that repository records of the archives studied are a

mix of full text and metadata. This needs to be considered when looking at population statistics. Minho's repository, on the one hand, contained almost exclusively full texts in 2005 (93%). However, the balance of full text to metadata may well be altered by the time it implements its CVs or online publication list service where many references may not link to open access full-text documents. This means that full text to metadata numbers reflect the aim of the repository and the services it provides. HAL discourages deposit without full text by setting a default to search for only full-text material in its portal. Southampton's institutional repository, on the other hand, though also pushing for full text acquisition, has a current research information system function. This consequently means that metadata-only records are sometimes inevitable, since publishers do not allow the storage of some material, or embargoes exist. It is clear that it is a challenge for the repository manager to achieve a balance between providing services that stimulate content acquisition such as CVs or that of a CRIS and the consequences that such services may have on full text aims and coverage.

Policy regarding the removal of content from the repository are defined in some of the selected cases. HAL has a specific policy to prohibit the deletion of any of its records in order to preserve its research. Soton, however, has a take-down policy which removes works which breaches copyright agreements. Other institutions will not remove the material but hide it, as with some Dutch repositories.

Recommendations

The following pointers for stimulating the population of repositories have been identified which relate to policy issues. For the complete list of guidelines see section 3.5.

1. Engage senior management to obtain high-level support. Use them to develop policy or services in order to fill gaps in repository stocks. Encourage them to establish mandates to deposit full text, and couple these with financial incentives to deposit where possible. If mandates are not yet viable, then implement incentives such as services to win hearts and minds. In the case of institutional repositories, strive for the repository to become a research management information tool or CRIS giving the repository a dual function. One-time registration of the recent academic record for internal evaluation can be combined with the repository function as storage and dissemination platform. Lastly, use senior management as a sounding board to further develop your repository and its policies in the future.
2. Use your local, regional, national and international networks by exchanging experiences and publicising work done for the purpose of policy and service development, personnel development and public relations. Exchange experiences with colleagues in a similar position, be they repository managers, policy makers, technical developers or communicators for cost-efficiency.
3. If national funding bodies mandate deposit use this argument when presenting the case for deposit to your authors.

3.4.2 Organisation

Organisational decisions and support are significant to the realisation of repository population goals. Drivers, institutional backing and governance structures will provide the policy, infrastructure and support mechanisms to enable or speed up efforts to be able to realise population ambitions. Understanding how research information workflows function and where they occur most efficiently will also contribute to achieving a more constant information flow. This section will address these issues using experiences gained by the six cases.

Factors for choices in organisational models

It is the scope and character of the service or repository which determines the organisational structures behind them. The Cream of Science service highlighted leading Dutch researchers and thus built its service upon 15 Dutch institutional repositories, including all universities. Connecting Africa is similarly a service layer on the data layer of several institutional repositories with a mission to bring African studies scholars and their research together into a broader international context via a portal. As of January 2007 it harvested 44 repositories for this purpose. In the case of HAL on the other hand, it promotes the central deposit of content using the same organisational model as arXiv.org. It does this partly to optimise efficiency in the aggregation of research output, federating efforts across disciplines and institutions in France. HAL claims that this model also serves to better control terminologies, author identifiers and metadata standards.

The institutional repositories such as CERN, Minho and Southampton each serve an entire campus and aggregate content into one repository. Southampton's ECS repository, however, manages its own output and OAI repository, which is then ultimately fed back into the central Soton institutional repository, as is data from other Southampton departmental archives and databases at present. CERN, with its international institutional profile, even goes so far as to aggregate thousands of records from non-CERN authors and puts CERN output in the context of an international digital library service for its particle physics community, thereby creating its own worldwide portal on particle physics.

Driving forces

All institutional repositories interviewed are aiming high for the 100% coverage of current research. The individual person drivers behind all services and digital repositories interviewed are dynamic and similarly ambitious. Repositories and services studied have been organised on various levels appropriate to the goals they fulfil. It is mainly the libraries who run the institutional repositories in this study, with the exception of the ECS repository at Southampton which is managed by a university department. Other federated repositories such as HAL or repository services such as Cream of Science or Connecting Africa have strong organisations behind them. CNRS, as the largest French multidisciplinary research organisation, took

the lead in establishing HAL with the CCSD running it.[34] Connecting Africa, a disciplinary portal, was established by the Africa Studies Centre, which is similarly an esteemed research organisation. This independent scientific institute undertaking social science research already had an established network of experts with a mission to 'promote a better understanding and insight into historical, current and future social developments in sub-Saharan Africa.... promoting the dissemination of knowledge and an understanding of African societies'.[35]

The SURF Foundation of the Netherlands, behind both Cream and CA, has done much for the institutional repository and open access movement both in the Netherlands and abroad. It is the national funder of network services and information and communication technology projects in the higher education community in the Netherlands and answers to the Dutch university boards. SURF managed the national repository network of institutions in the form of the DARE community between 2003 and 2006. It is this, combined with milestones and funding for projects instigated by such a powerful organisation, which has been a significant factor in seeing successes in institutional repository population and service development in the Netherlands. This leads to the conclusion that large organisations with influence are in a good position to take on ambitious cross-institutional projects. Larger federated initiatives need influential drivers to become established, adopted and populated.

High-level support

High-level management support cannot be underestimated. The cases studied in this chapter have considered this crucial to establishing policies that can contribute to repository development, take-up and population. The libraries interviewed have consulted with and advocated their aims to high-level managers including rectors, university boards and research faculty heads. In some cases, cultural changes have been made within the organisation by establishing mandates to deposit electronic copies of academic output. High-level support by leading individuals is clearly of significance for institutional buy-in to a repository or service. The Minho Library is clearly indebted to its university's rector, who both supported the establishment of policy proposals and actively contributed to them with his own ideas, including the introduction of financial incentives to deposit. Southampton involved its deans of research and faculty management in the establishment of its repository as a result of the pilot TARDIS project. SURF, as leaders of the Cream of Science project and funders of Connecting Africa, has a high-level university board behind it when making decisions of national significance and influence. One layer below this, and above the operational level, is the steering group of the Dutch library directors, which decided on policy issues surrounding Cream. This further guaranteed that participating institutions were dedicated to meeting ambitious goals.

HAL was established as a national product. Eighty-six of the council of rectors of France's universities committed themselves in writing in 2006

to collaborating with HAL in the future. HAL hopes to extend its content considerably with this agreement.

The Connecting Africa team sees the backing of the African Studies Centre (ASC) director, as lead of Connecting Africa, as a critical success factor in the establishment of the service, its maintenance and further development. Its maintenance has been made part of the ASC's annual work and strategic plan for 2005-2008, for example. The establishment of Connecting Africa has been an important stimulus for the instigation of such an internal policy and hopes to contribute to higher content deployment figures. CERN's director-general has also been supportive in endorsing CERN's e-deposit mandate, and publicly and internally addresses the importance of open access to CERN at scholarly publishing events, as at the European Commission in February 2007. Such actions can further stimulate internal support for open access and encourage or even press more researchers to deposit.

Governance

Various governance structures were identifies between cases. The ECS repository is governed by the school's management committee, and reporting takes place three times a year. The Soton repository, however, is guided by the Southampton institutional repository EPrint steering group chaired by the university librarian, which meets bi-weekly. This is where strategy and policy issues are discussed, as is liaison with other national repository projects. The Minho Library manages its own repository and only reports to the rector, although establishing a strategic board is under discussion.

In the case of HAL, various committees advise: there is a scientific and technical board as well as a strategic committee with representatives from some of France's leading research organisations, and the CCSD president.

Connecting Africa and Cream of Science both had project boards, Connecting Africa's project board contained both scientific and library staff whereas Cream used its DARE steering group largely of library directors. The challenge with projects is that once they come to an end, so do their governing boards and other models need to be found for sustainability. For example, the Cream of Science service has handed over its service to the Royal Netherlands Academy of Arts and Sciences, which maintains a comprehensive Dutch research database. Running Cream is now part of their scope and responsibility.[36]

Autonomy for research departments

From the cases studied, evidence has shown that deposit organised on departmental or school levels seem to reap the best results. The Minho Library director, Eloy Rodrigues, points out that establishing a policy which allows some autonomy to the researchers involved has been a critical success factor for populating his repository and partly explains wide departmental take-up. Allowing communities to exercise some influence on the content to be aggregated and profiled engages them in the decision-making process and accommodates different community needs. Minho also has its

own departmental repository coordinators in place to support its researchers. This organisational model can then also cause research departments to consider deposits a responsibility and product of their own thereby further motivating population. CERN has similarly organised this decentrally by mainly making departments responsible for content deployment. It is the ECS repository at Southampton which was even established and run by a research and teaching school, specifically the school of electronics and computer science, seeing it as in their own interest to aggregate, organise and disseminate their content more effectively.

Analysis of workflows

The successes of collection development for both services and repositories is dependent on the knowledge of workflows so as to ensure the provision of simple depositing infrastructures. This was mentioned as a critical success factor by both CERN and Southampton. Connecting Africa explicitly mentioned analysing the publishing workflows of Africa Studies Centre (the initiating institution) researchers in order to better organise the internal scientific information management processes. This was also the beginning of the Connecting Africa service. Management information and the analysis of where this information is stored and how this is organised, whether this be in research departments or in a CRIS, for example, are therefore essential to understand how to more efficiently organise and aggregate material for a repository. For more information on research evaluation, see section 3.4.4. Services.

Recommendations

The following pointers to stimulate the population of repositories have been identified which relate to organisational issues. For the complete list of pointers see section 3.5.

1. If you seek to develop a regional or disciplinary service to help populate your repository, choose a prominent partner with influence, preferably well recognised by the library or research community.
2. Consider the distributed organisation of academic output within the institution when planning population. Consider the structure of your organisation and adapt to it.
3. When organising your repository, strive to give some autonomy to the research community, giving them the responsibility to maintain their output with library support, encouraging them to feel like the owners of their own output.
4. Strive for cost-effectiveness by analysing work processes to synergise with CRIS, research management information departments, disciplinary portals, or funding agency mandate practices. Aim for departmental and/or researcher deposit and not library deposit for long-term sustainability. However, as a library, do consider repository investment for content acquisition and invest time to provide a good demonstrator to encourage self-deposit in the future.

3.4.3 Mechanisms and influential factors for populating repositories

A number of factors will influence the population of a digital repository. Collection development decisions and content deployment methods will form the scope of the archive. These are heavily influenced by the disciplines a repository serves. Researchers may come from traditions of self-archiving or be part of a community which publishes research results in conference proceedings and monographs rather than journal articles. This affects the collection profile and ensuing numbers of full-text files in the repository. It is important to be aware of researchers' motivations regarding contributing (or not), which may be discipline-specific or generic. Awareness of these factors can help form plans to fill gaps in repository stocks. This section will explore these issues based on input from the cases studied.

The importance of the discipline

Successes of a repository are not to be measured by full text file numbers alone, this is a quick but incomplete measure of evaluation. Successes need to be balanced against the aims of the repository or service: is it providing improved access to open access research output? Is the material aggregated to serve for an evaluation exercise, or will the repository also serve the researcher in information retrieval and discovery? In the case of either the former or the latter, the researcher as information provider must remain the focal point. Repository managers interviewed have stressed the importance of involving research communities in the decision-making process of the academic output to be gathered. In the case that an information service is developed for a specific research community, for example, relevant material needs to be available to support research. For example, Connecting Africa deposits are split up mainly between full-text files and images which are of significance to ethnological studies and cultural anthropology (27% and 72.5%, respectively). This fact should dispel the notion that repositories are only successful when numbers are measured against material of a textual nature. This hypothesis will be further strengthened when more teaching and learning materials and especially data sets populate repositories in the future – a goal expressed by many interviewed.

Population results are very much dependent on the authors' discipline and self-archiving traditions. CERN excels with its close to 80% coverage of academic output serving a community which has been self-archiving since the early 1990s. However, CERN has other challenges to contend with where authors prefer to deposit material to an international disciplinary repository which has been serving its needs for years. In the case of serving a community already accustomed to self-archiving and familiar with data deposit centres external to the home institution, mechanisms to reclaim material need to be developed for collection management. CERN has done this by harvesting material from a large number of data providers. CERN even

states that without such actions that it would have 'dramatically failed' in achieving its collection development goals.

At Minho some research centres deposit 15 to 20% of their output, others like biology, bio-engineering and civil engineering deposit between 80 to 90%. This is despite the fact that 75% of all academic staff contributed to the institutional repository in 2005, the year of the mandate, and an 87% take-up in establishing repository communities by its research departments. This discrepancy again shows the significance of the discipline in the population of a repository. Achieving 100% current content will be dependent on groups of research. Self-archiving traditions, scientific and management interests in collaborating, and the ability to deposit in an environment which is currently controlled by the publishing market are all aspects which will influence current content goals. For example, some disciplines are more dependent on monographs for disseminating research output. In such cases, a 100% current content aim is probably too far-reaching. Eighty-seven per cent of the 30 research communities at Minho show a commitment to collaborate, but campus-wide annual academic output figures are still not reaching the 50% mark. Here, one can hypothesise that the intention of management is to collaborate, but that the researchers do not wish to or are not able to comply for some of the aforementioned reasons. This gives further reasons to home in on the needs of the individual researcher in addition to those of senior management.

What effect disciplinary take-up has on the actual deposit or growth statistics of the repositories can be seen by looking at deposit statistics with relation to annual academic output. This peaks at 80% at CERN, with Minho with a 40% average.[37] These are success stories and are not representative of figures elsewhere, but even with all departments formally committed to deposit with dedicated local coordinators as at Minho, the 100% figure has not yet been obtained.

Researcher take-up

The early adopters listed here relate to disciplines at the institutions with repositories. At Minho, a university which addresses the arts and humanities, social sciences, health, life sciences and engineering, early adopters were from the areas of bio-engineering, engineering, systems information and management research. Bio-engineers at Minho in 2006 deposited approximately 90% of their research output. Soton saw early adopters from the schools of electronics and computer science, oceanography and earth sciences, optoelectronics research, and engineering sciences and education from a university with faculties of engineering, science and mathematics, a faculty of law, arts and social sciences, and a faculty of medicine, health and life sciences. At HAL, which covers all disciplines, mathematics and social sciences and the humanities (with an emphasis on the humanities), were their early adopters. This was partly as a result of the sub-portal developed for these areas via http://halshs.archives-ouvertes.fr. This shows that developing views on specific disciplines can encourage the deposit from disci-

plines which are normally more reluctant to contribute elsewhere, such as the humanities.

In the interest of content deployment, similar communities can be targeted which can then stimulate other communities to deposit. Disciplines with little to non-existent self-archiving traditions that are less dependent on journal articles and conference proceedings, authors of books or editors with economic interests in maintaining their own society journals see discouraged interest. If 100% coverage of current content is still the aim for many, the more challenging disciplines such as law, the humanities and the social sciences with such views need to be addressed more specifically as do their publishers. For more information, see section 3.4.6. legal issues.

Researchers collaborate and populate repositories for a number of reasons, according to the repository managers interviewed. These range from reasons of obligation due to a mandate to deposit in place as mentioned by Southampton, CERN and Minho, to increasing visibility and impact, to accessing new material, to supporting the principles of open access. The first reason to deposit listed by Minho was to increase visibility and impact through the further exposure and dissemination of academic output. This aspect was mentioned by all those interviewed. This fact, and considering its position in listings, leads to the conclusion that this is one of the most important factors for contributing work to repositories. CERN has evidence of this potential increase in visibility, and goes as far as to say that on the analysis of user logs, 70% of those who consult its repository content come from outside of CERN.

The development of services to answer author-specific problems have contributed to raising the repository profile. As a result, HAL mentioned its services as a motivation to deploy content. All cases who provide automated publication lists as a service based on their repositories for example indicate this as a reason for researcher take-up. Southampton mentions utilising such lists for web pages and RSS feeds as incentives to collaborate. Access to newly digitised material was explicitly mentioned by Minho and Connecting Africa as contributing factors. Support for and conviction of the future in the open access movement was mentioned by Cream of Science and HAL leaders as reasons for content deployment by its researchers. Other additional reasons mentioned were the Research Assessment Exercise and the organisation of research articles by Soton, for reasons of preservation at Minho and due to the fact that much work is done on behalf of the researcher resulting in little effort. In the case of Connecting Africa, the researchers' affinity with the lead institution behind the service was also felt as a motivation to collaborate. Due to the nature of the Cream Science, which showcases leading Dutch researchers, prestige was indicated as a drive to cooperate by Cream's instigators.

Reasons for researchers not contributing or not having contributed in the past mainly revolve around the fact that the benefits and added value offered by institutional repositories and their services were sometimes unclear to them, which the above-mentioned activities are trying to resolve.

Population can also be hindered by some authors who are reluctant to make too many versions of similar works available via open access which might highlight repetitive work worldwide; this has also been reported as a problem by CERN authors. For more information on such inhibiting factors, see the DRIVER website www.driver-community.eu

Content deployment

In the cases studied, deployment takes place both centrally and decentrally by author, representative or library staff. There are various reasons for this. The HAL model was initially designed, like arXiv.org, for researchers to directly deposit material into one central archive claiming to thereby more directly serve the needs of the researcher. This still takes place in addition to some administrative or information support staff from institutional libraries – as is the case for the CNRS physics community – who assist the researcher in content deployment.

HAL has a standard interface in situ, but has also designed various institutional or disciplinary interfaces to its archive to encourage further population by both individuals and groups of researchers.[39] It also accommodates for the integration of institutional and laboratory local archive content through web services.

Deposit is generally organised on a departmental or research group level by most digital repositories interviewed. Minho University has established, partly due to the DSpace community repository organisational model, and partly following the university research structure, 26 separate repository research communities out of a total of 30 possible ones.[38] Research departments all follow their own policies for depositing certain types of academic output significant to them. Communities have local coordinators who are responsible for providing basic support to their researchers having been provided with various manuals and documentation by the library to support them. A library repository helpdesk helps with questions which cannot be addressed by the research centre coordinators. CERN input is similarly organised on a departmental level where departmental secretaries administer preprint entry and some other publications.

Soton material is deposited via departmental databases (as of June 2007 approximately 25% of the total) or repositories be they bibliographic, full text, open access or not. The breadth of sources partly has to do with the historical organisation of research output as well as with the management structure and autonomy of schools at the university. ECS has its own open access repository which is maintained by its own school, although this is the exception to the rule at Soton. This decentralised data is then fed into the Soton institutional repository. However, Soton is also seeing increased direct deposit to its central repository; some schools are abandoning their own systems for a library-run one.

The institutional repositories studied have the policy that individuals or research departments deposit in their institutional archives, and that the library does not do so on their behalf for reasons of sustainability. Soton and Minho therefore also mention the importance of a simple deposit sys-

tem. In the cases of Connecting Africa and the Cream of Science the libraries fed the services which have played a significant role in the deployment of metadata and full text. In these cases, libraries have indeed entered data on behalf of their authors in many cases in order to achieve ambitious targets and to gain content rapidly.

Content ingestion

Harvesting

Harvesting can be used to add to a repository's content stock. Soton and Minho do not harvest at present, and rely solely on departmental or individual deposits from researchers and their administrators. However, HAL does harvest parts of arXiv.org as an exception to its rule, as does CERN. This is an essential activity for CERN for further populating its repository. 'Had we not aggregated material from outside we would have failed dramatically', says Jens Vigen, institutional repository head. CERN harvests material from a range of approximately 90 external resources to further obtain complete coverage of its recent and current academic output, contending with researchers who deposit elsewhere, and not with the institutional repository. Entire archives to subsets are harvested from once a year to once a day, and by agreement as long as existing services do not suffer as a result. This includes full text harvesting. This is a significant contributing factor to reaching CERN's 79% of its academic output where approximately 30% is gained by this means.

International disciplinary services which are based on repository content harvest from others to bring together research for improved access such as Connecting Africa or the national Cream of Science showcase.[40] Connecting Africa, for example, harvests entire repositories and then filters the relevant African studies content by utilising tools which select data based on subject lists. Filtering harvested material thereby extends the scope of its service making it of greater international interest to the research community and stimulating further deposit. This filtering of harvested material is a way of cost-effectively populating a service, stimulating the reciprocal population of institutional repositories which contributes to the service.

Cream of Science and Connecting Africa services mainly rely on the content of institutional repositories for the provision of their content. However, HAL, which collects the academic output of French research, does not generally harvest. HAL claims that through direct deposit and web services it can 1) more directly serve depositors, 2) the level of data reprocessing is lower with such a central system and 3) the control of data deposit is more efficient. For more information on this, see the HAL case study write-up.

SOAP web-based services

HAL deploys SOAP web services where material is exported to HAL from current archives through remote submission.[41] Connecting Africa uses a mixture of web services and harvesting to disseminate and retrieve its con-

tent. Connecting Africa even mentioned SOAP web services as a critical success factor for maintaining some of its content more efficiently.

The influence of collection development decisions

Current content versus retro-digitisation

All repositories interviewed are aiming for 100% coverage of their recent or current research. This regards both metadata and full text. Soton, CERN and the ASC behind Connecting Africa also provide information for internal evaluation with a CRIS-type function through their institutional repositories, which supports this mission.

To encourage current content deployment, Minho established a policy to ensure the deposit of quality and current content across the campus where authors were reimbursed with higher research points were the articles they posted recent ones. There is no HAL programme for acquiring current content although HAL reports that evidence has shown that this is generally the first to be deposited.

However, although services strive to keep material up-to-date, this is not always possible as the self-archiving policies of publishers are either unknown or embargoes rest on the publications as mentioned by Connecting Africa. Publishers also stand in the way of providing current content, claims Soton, due to the national agreement between libraries and publishers to only store Digital Object Identifiers rather than full texts in institutional repositories for the RAE. This has hampered the population of full text current content, although evidence has shown growth in full-text deposits despite this.

There is no policy to retro-digitise at any of the repositories interviewed. Soton does not generally carry out retro-digitisation due to its current content focus although it will carry out the digitisation of theses were the need to arise. Retro-digitisation campaigns do exist to put theses online in France, for example, although this is not a HAL initiative but rather one of research institutions. Little retro-digitisation is carried out at Minho, although it has carried out some to encourage certain newcomers to deposit in 2005 when certain important research material was scanned to increase digital access. Soton similarly adapts to the needs of its users by including important historical output or publication list references, which means that the repository is also being populated with older (often purely metadata) material.

It was the Waaijers pre-digitisation era policy, which permitted the digitisation of pre-1998 journal articles, which had a positive effect on the population of all institutional repositories in the Netherlands. Larger retro-digitisation programmes as part of Cream for example have produced approximately 20,000 new documents, making more material available online and accessible via services such as Cream and Connecting Africa. However, during this period this policy has resulted in the population of repositories with less emphasis on current material, where priorities lay on numbers rather than on currency with critical mass as the main focus.

Dutch institutional repositories are now therefore also focusing on increasing the aggregation of current content. Retro-digitisation efforts will clearly contribute to repository numbers, however, it should be pointed out that this is not always possible due to national IPR laws.

Content type

Population figures are affected by the content scope of the repository. Academic output is aggregated by all repositories and services studied although some collect more varied types of information than others. All aggregate journal articles, and most also aggregate books, chapters, proceedings and working papers. Minho, however, concentrates on providing journal articles (41%) and conference proceedings (40%).[42]

Theses and dissertations are aggregated by Soton, Minho, Cream and Connecting Africa. Soton's School of Oceanography's theses are being electronically deposited in the repository. HAL reports not storing dissertations. In fact HAL has a system of virtual document collections, which consequently form types of publication bundles. Images, audio or video files may only be submitted as annexes to other documents for example, creating an enriched publication.

As Soton points out, the scope of the content type is discipline-dependent. For example, ECS and its computer scientists aggregate journal articles and conference papers which are the main means of publication in the computer science area. Midwifery and nursing however highly value conference proceedings, and medicine regard journal articles more highly. CA has a large collection of image files as this is important material for ethnographic and anthropological research. Disciplinary needs are therefore reflected in the collection profile of the repository particularly when research departments can determine what can be deposited as is the case at Southampton or Minho. Certain cases concentrate on building on their content stocks with particular types of material. This is either by continuing building on present collections, such as HAL aggregating more scientific articles and theses, or else focusing on brand new types of research output for institutional repositories such as primary data sets which is the case for CERN, Minho, Cream, Soton and possibly HAL in the future. CA will even go beyond academic output and will aggregate content from policy makers, NGOs and journalists, thereby bringing policy and research closer together in the future.

Less frequently repository-stored data types at present are audio files, music, and images. CERN also collects teaching and learning materials, which is something for the future for Cream and Minho. Cream also provides access to inaugural lectures, speeches and newspaper articles of its leading authors. Lectures and software are also stored by HAL. Soton is open to the storage of new media types such as streaming videos.

Versions

Population numbers are surely influenced by the aggregation of multiple versions of one and the same paper or of similar versions. All cases inter-

viewed store various versions of publications. Soton aggregates any versions of material from the schools which are classed as academic output, although post-prints and publisher versions of journal articles are preferred and reflected in Soton policy. Cream, CERN and CA for example make various versions of papers available such as working papers, post-prints and publisher PDFs. HAL authors may also post various versions of a publication but under the understanding that deletions of any full-text articles are forbidden for technical reasons as well as for preserving the scientific record. In addition, CERN is making efforts to collect PDFs from publishers which allow the self-archiving of publisher versions to boost full text population figures.

Certain versions such as publisher PDFs or post-prints are unavailable via the repositories due to copyright restrictions. In the Netherlands, and through Cream of Science, it is known that some of these versions are indeed made available either due to the pre-1998 copyright ruling, or as a result of requests by more liberal authors or library policies. Interesting author behaviour at CERN can be observed concerning the deposit of certain versions of papers which are very similar in nature. In some cases, some authors are reluctant to deposit any material in fear of making repetitive work more transparent through open access. Further visibility through the institutional repository of this fact is a motivation not to contribute to it therefore and can therefore hamper repository targets.

Recommendations

The following guidelines to stimulate the population of repositories have been identified which relate to mechanisms and influential factors for populating repositories;

1. Know your research community, and address their disciplinary needs specifically by speaking to the author as part of a specific disciplinary group with specific needs. Differentiate between young and old authors regarding deposit considering their differing work processes, challenges and motivations to publish.
2. Ensure that the repository and its services serve the researcher, answering real needs and resolving author and reader problems.
3. Make collection development choices which reflect the academic output of your disciplines. Focus on the challenges in unlocking that material, as self-archiving traditions and possibilities differ widely across disciplines.
4. Be innovative as to how you acquire your content. If you are at first unsuccessful and your researchers deposit elsewhere, identify which archives are important places of deposit for the researcher of a particular discipline. Make agreements with those sources to either harvest metadata or full text, and monitor that data.
5. Be cost-efficient by harvesting content from outside your repository to further acquire missing content and use SOAP services to update content.

3.4.4 Services

The cases studied point out that inertia and aversion to open access activities function as inhibiting factors for populating repositories. Researchers do not want yet another administrative task to contend with and to be under further control by institutional management. Repository managers interviewed see the provision of real benefits to the research process through the development of added value services as the way to challenge and break these preconceptions. Services are being developed as incentives to maintain or stimulate the further population of repositories. Such services will furthermore uphold and provide further digital library services. The palette of services provided by the cases and the issues they address will inspire those looking for means to encourage further content deployment.

Services are built upon the data layer of either individual repositories under study here, or built upon multiple repositories feeding the Cream of Science or Connecting Africa services here. Added value services have either been built for researchers, research management or for university management based on issues of common concern. A palette of approximately 30 services have been developed by the cases studied. Almost all institutional repositories studied are currently active with technical developments, which has helped to implement solutions and services at a faster speed than many could hope for. For a comprehensive list of services, see the case studies on the Driver website www.driver-community.eu

Increase visibility to institutional research results worldwide

The open access movement and researchers alike are interested in increasing the impact of research. Repository leaders are keen to increase access to research output, and make repository contents visible via open access worldwide, and thus more easily useable. The cases studied here have all been involved in trying to ensure the increased global visibility of research results. Efforts have also been made by several of the cases studied to make repository content available where the research readership is to be found. This is done by making material available to specific subject-specific search services such as SLAC SPIRES-HEP as CERN does, or by pushing material out to disciplinary archives such as arXiv.org, RePec or PubMed Central by HAL.[43] This means that one-time deposit results in the visibility of data in multiple information services. More generic cross-disciplinary services are also targeted be they national, for example, the UK's portal for key resources for education and research INTUTE as targeted by Soton, or international search services engines such as OAIster and BASE as targeted by Minho.[44] Google or Google Scholar are mentioned in most cases as places where institutional repository material is further disseminated.

Not only is the visibility of research output important in these search engines, but certain cases interviewed have made efforts to optimise the positioning of their material within them. Soton has analysed Google rankings in order to determine whether it can influence the place of its material in future; Minho has done the same for Google Scholar and hopes to ana-

lyse more similar services to better target the visibility of its research. CERN has agreed on metadata specifications with Google Scholar whereas HAL plans to insert its Dublin Core metadata into the html meta-field of each HAL publication page to ensure the better information retrieval by such services.

In cases where services are dependent on institutional repositories for their content such as Cream of Science, where the dissemination of institutional repository content is seen as the responsibility of the institution, Cream has checked that sub-sets of its content are available via international subject-specific services such as Connecting Africa. Yahoo, OAIster and Google Scholar bots also pick up Cream content. Ideally, Cream's content should be part of a broader international knowledge base of academic research results. This explains its initiator's involvement (SURF) in the DRIVER gateway also ensuring that the Cream set of Dutch research is visible and searchable in a broader international and interdisciplinary context.

Information discovery and retrieval

Information discovery and retrieval on the web is generally a given in today's research practices. For this reason, all cases have implemented search and browse facilities on top of institutional repository content or have developed larger portals on top of multiple repositories like Connecting Africa or Cream of Science for rapid access to research references and full text. Minho allows cross searching of its DSpace communities as well as the searching of individual departmental content. The Minho Library encourages its new students to use the repository as an information resource. CERN, however, goes further in providing its institutional repository content in the context of worldwide content related to the subject-community it serves. CERN acquires this content by harvesting a mass of worldwide content from approximately 90 databases worldwide, including arXiv.org, both to reclaim some of its own academic output and to provide a CERN particle physics digital library service. This so-called 'CDS database' search service therefore enables the information discovery and retrieval of world particle physics content. In addition, CERN has developed services around this portal such as hyper-linking citations, other authors who consulted document x, also consulted document y. This is clearly easier to realise in serving one subject community but a great feat for an institutional repository with a broad spectrum of different subject communities to answer to.

The HAL interdisciplinary portal on French research can be searched by author, institution or virtual collections based on subject domains. Collaborations between institutions can also be identified by searching for publications by keyword thereby highlighting networks be they institutional or people. The Cream of Science multidisciplinary portal showcases leading Dutch researchers and their work, and its search system allows searching by author name, organisation and year as well as by keyword (as part of the title). Material can be browsed by institution, author and discipline. The Connecting Africa portal on the other hand is subject-specific in nature

providing access to the names and work of Africanists throughout Europe. Publications and images can be searched by title, author and keyword or via a simple search box through which experts can also be found.

Additional services have also been developed to encourage users to return to services or to raise awareness on new content added to the system. Connecting Africa for example aims to attract users to return to its portal by having a 'pick of the month' resource feature which appears on the first page, which is selected by the ASC digital librarian. With the same incentive, HAL features its last five deposits and publicises its growth and deposit statistics via graphics on its home page. RSS feeds are also available via Soton, Minho and HAL repositories to push back new content to end users.

Save time on administrative tasks

Library experience worldwide has shown that researchers fear losing valuable research time to extra administrative tasks. It is therefore important to identify services which save the researcher time on administrative tasks rather than adding time with an additional one in posting material to a repository. The creation of automated publication lists which link to repository full texts seems to be a service which has been welcomed by researchers, who are otherwise manually maintaining lists of publications, often with few links to full text. Such automated publication lists services are offered in most cases studied, and seem to be becoming a standard service for many institutional repositories. This service is offered by Soton, and CERN, as well as by repository services such as Cream of Science and Connecting Africa. Minho will generate publication lists for CVs in 2007/8.

HAL also allows the researchers to select and sort publication references based on personal preferences for personal home or lab web pages. Publications can be listed by type or lab, which means that researchers, groups and institutions can showcase selected research output. Data can also be exported, as is the case with CERN. Soton will also in future export publication lists and full texts for grant applications and to update CVs.

Promoting the one-time deposit of content for utilisation by the CRIS and repository is also a means to save time on administrative tasks. Several institutional repositories have a dual role as a type of CRIS and repository. Soton points out that the Research Assessment Exercise (RAE) and the Soton institutional repository's role in serving it have acquired institutional support for the repository. Population figures confirm this fact. Such services could well enhance researcher collaboration and thereby contribute to the population of repositories. One can therefore conclude that a motivation to deposit could be one-time repository deposit resulting in multiple administrative outputs such as publication reference lists, CVs, CRIS outputs and posts to external information services for further dissemination and visibility.

Research assessment

An obligatory administrative task for many European researchers is recording academic output for evaluation purposes. Soton sees the UK's Research

Assessment Exercise (RAE), which obliges all UK higher education institutions to record their academic output over a certain period, as a critical success factor for populating its repository. Southampton's institutional repository has obtained the function to aggregate the content necessary for RAE evaluation. Considering that the results of the RAE are essential to Southampton's future research funding, Soton schools are forced to comply to deliver references if not full text for this exercise. As a result, institutional buy-in is acquired for the repository which further guarantees current content deployment from all disciplines. The institutional repositories of CERN and the ASC for Connecting Africa have similar current research information system (CRIS) roles for their organisations. Minho hopes to run the university CRIS, or at the very least link it with its institutional repository, generating publication lists and CVs from the system. Not only is this useful for university management evaluation as a whole, but also for department evaluations. Research communities, for example bio-engineers in Minho, already use institutional repository content and its full texts to evaluate their own output, resulting in a surge in deposit of the repository shortly before an internal departmental evaluation.

Market research results

Southampton, and particularly ECS and its management, sees the institutional repository as a means to market new research output. New material deposited is highlighted on web pages and on departmental plasma screens, for example, and departments can now track which papers are heavily used, for example, on the delivery of download statistics. ECS believes that publishing usage statistics has had a positive effect on population statistics and on stimulating others to contribute.

Disciplinary views on research results extracted from repositories can be used for both research assessment and marketing purposes to encourage population. HAL provides windows on labs or subject domains to show academic output and collaborations in certain areas. Disciplinary portal interfaces have therefore been developed for the humanities and social sciences at http://halshs.archives-ouverts.fr. Institutional repositories at Soton and Minho have similar services where departmental views on research can be seen, and Minho research department content can be viewed by title, author or subject. This allows departments to profile and increase access to their work both for the improved internal accreditation of work as well as for better external visibility.

Provide further insights on research impact

Southampton provides usage statistics to research groups and individuals via the web. This demonstrates the impact that research online and open access can have on the use of the individual's research results. This is also actively being used by researchers and their groups at Soton. Download statistics are also delivered to authors by HAL. Minho will implement such a service in 2007. This is a method to prove the value to the researcher in

depositing material into repositories by showing publication usage figures of a broad scope which are generally not obtainable elsewhere.

Preserve access to research

Preserving access to research output is an argument which researchers embrace as a motivation for depositing material in a repository. On the most basic level, all archives serve as back-ups for research results. However, as far as long-term preservation and access to files and their re-use in the mid to long term is concerned, most cases studied see this as the responsibility of national libraries or national initiatives. Organisations can prepare for this, as Soton is doing, by creating a catalogue of data formats. ECS intends on carrying out format migration when necessary and CERN is likewise prepared to convert to new formats. HAL is currently converting documents into XML or PDF/A formats to safeguard long-term preservation. Safe copies are ensured in a secure storage environment together with highly secured research data at a supercomputing centre in the case of HAL.

Preservation is one of HAL's main missions and Minho mentions it as a key reason for users deploying content, so this will need to be addressed further than present capacities allow in order to fulfil expectations with long-term preservation efforts often not in the scope of such entities. By the nature of HAL's organisation, scope and mission, HAL claims to be in a better position to deal with this than individual organisations. Institutional repositories may alternatively, as several have already indicated, need to turn to national libraries to tackle this complex issue, which brings me to the Netherlands.

In the case of the Netherlands, the Dutch Royal Library now stores all Dutch institutional repository content perpetually, having drawn up contracts with each institution in 2006. This means that Cream of Science full texts and selected parts of Connecting Africa are securely stored and accessible in perpetuity. This is one of the first of such initiatives of its kind worldwide.

Added value services in summary

It is HAL which provides the most services of all the case studies: thirteen as of February 2007. These have been mainly developed to support the researcher in populating the repositories in question. For more information on these services, see the HAL case study write-up on the Driver website, www.driver-support.eu. The top three services which have been implemented most in the different cases studied are search and browse services, automated publication lists and the electronic dissemination of information to external information services. Seventeen services are unique to the other cases.[45]

Service use

Interviewees often had little insight into the use of the services which they had developed when asked – be they disciplinary ones based on a number

of repositories such as Connecting Africa – or sub-services of institutional repositories such as publication lists. However, going back to the motivation for the development of most services, which is to populate repositories, evidence has shown in Cream of Science, for example, that authors are willing to contribute their content to institutional repositories for the purpose of such services and all cases have seen similar experiences. Changing the reader's information retrieval habits and sources of research is another challenge in itself, and perhaps one which will be reduced come the day when repositories and their services are filled with a critical mass of quality content which the user then considers as a relevant port of call.

Recommendations

The following pointers to stimulate the population of repositories have been identified which relate to services. For the complete list see section 3.5.

1. Provide added value services which are flexible and adaptable to save the researcher time on non-research activities.
   Seek to ensure that a demonstrator is in place and strive to ensure that as little time and effort is needed to deposit material. Examples of services are CVs or automated publication lists, search and browse facilities to allow as much cross-interrogation as possible, with the possibility to discover not only new research, but networks be they institutional or people. Push out information from your repository to disciplinary services on behalf of your researchers. Use RSS feeds to feed back new material that enters the system to various disciplinary groups. Retro-digitise older material; seek to preserve the academic record, convert formats, and seek to implement long-term programmes in collaboration with others. Lastly, monitor the use of these services.
2. Showcase your efforts and achievements by marketing your research results by publicising recent repository additions by research area/department feeding them back to authors and/or research department. Install log analysers of download and upload statistics and feed back this data to depositors and their research groups. Promote the dissemination of this information via departmental websites, individual web pages, etc.
3. Push out your content to the world research community to show your commitment to increasing the impact of your researchers' work. Do this by liaising with your researchers to identify which sources are of significance. Ensure that you get your repository indexed or harvested by such information services and seek to optimise the positioning of your material in these sources.
4. Take on an active role in improving on information retrieval and discovery by contributing repository content to regional or international services of scientific significance, feeding back results to your researchers. Target information services such as Google, Google Scholar, as well as disciplinary ones.

3.4.5 Advocacy and communication

Resistance to repository activities in the author community is known, as reported by Davis and all cases studied here.[46] Advocacy aims to inform, dispel fears and provide the necessary support to bring about the population of a repository or service. Advocacy efforts need to engage and address all stakeholders on all levels, for example, from management to researcher (reader, expert and author) to administrator. Clear benefits need to be in place which address real researcher needs and problems, and the institutional repository needs to have the added value services in place to answer them. Ambitions are high in the cases interviewed (aiming for 100% coverage of academic output) and population is still a challenge for them despite successes. This section addresses advocacy issues focused on the research community.

Advocating open access and the repository to management and researchers
Two general groups of focus can be identified from the study: those who need to be bought into the concept of open access and institutional repository deposit, and others who need support whilst depositing where relations need to be maintained.

Staff resources for advocacy do not necessarily reflect success in population figures, however, they are important to be able to realise more challenging and time-consuming advocacy activities. Increases in numbers as a result of those efforts are evident. Minho on the one hand has a full-time, dedicated repository liaison officer who has developed a communication plan, liaises with library and faculty representatives, and provides much support, and has been mentioned as a critical success factor for Minho achievements second to its mandate. On the other hand, CERN, reaching close to the 80% mark of academic output, confesses to not invest much in advocacy. Library staff at CERN use opportunities to talk with authors on a one-to-one basis when opportunities arise to obtain missing work for its institutional repository. On a level between these two options, at Southampton, discipline-specific library academic liaison librarians advocate the institutional repository once the initial advocacy activities has been carried out by senior library staff at senior management levels. They inform researchers of the benefits of deposit such as further research impact and potentials for collaboration and demonstrate specific services which can assist them in their research such as bibliographic streams for web pages. This is given in the form of presentations and seminars. Similar messages are passed on by HAL in road shows where D. Charnay, as operational head, goes to significant research organisations in France to encourage deposit.

Advocacy efforts can be either author-centred or institution-centred depending on the goal of advocacy. In the case of projects which are dependent on the commitment of institutions to reach a large scope of authors, although the author is key, the organisation that will support the researcher in delivering content is also crucial. For example, in the case of the national HAL service, it is the research institution whose commitment and goal to

turn will into action that will help further achieve population HAL's aims. It is the organisation which often provide the support structure to assist deposit, and HAL road shows are geared towards them and their researchers therefore.

Wendy White, head of the Soton institutional repository, and her Southampton institutional repository team reported seeing advocacy as invaluable in gaining take-up across campus as did Minho where schools and senior management learned of the potential of open access.

Addressing people on all levels involved in the depositing process is also important, indicated White of Soton, be they managers, research groups, researchers or secretaries. This is similarly done at Minho. The diversity of needs can probably be best met through individual consultation though this is not scalable. This is CERN's stance at the moment, though probably a choice for reasons of practicality due to lack of resources for this area at present.

Advocating a service of international subject-specific significance to the research community is also important to stimulate its further population as well as to gain users of the service. This was the case for CERN, which needs to persuade its users to deposit with its own repository as opposed to with other data providers on the one hand, and the case for CA on the other presenting a new international portal to the African studies community. CA did this by presenting at least four national and international African studies events, one even in the United States.

CA also pointed out the importance of choosing the right time to advocate an institutional repository or service. CA feared losing its researchers were it to involve them prematurely without the necessary infrastructure and support in place, thus waited until its basic portal was available where deposit structures were available. Southampton also emphasised the importance of having a demonstrator in place. Cream on the other hand, had little to demonstrate but an idea at the outset although the repository infrastructure was in place.

Public relations material

Cases have developed various communication tools and PR material ranging from leaflets to posters to a desk calendar featuring international events of interest by Connecting Africa. Wendy White, repository manager of Southampton, does not believe that PR material is the best way of tending to various researcher needs and prefers therefore the personal approach. Her organisation does nevertheless use posters and flyers to share Southampton experiences with the information professional community. Minho seems to have the broadest scope of activities of the institutional repositories studied, including the above as well as a brochure to mark the RepositoriUM's anniversary and flash movies which show impact indicators of its use with download figures and visitor statistics. Flash upload statistics and visuals on the type of data contained can also be found on the HAL website.

Websites or portals exist for all. The Minho website for example, was developed to support communities in the self-archiving process, with search and browse screens, help information on self-archiving and copyright, as well as FAQs. Most repositories interviewed provide similar sites. Soton and Minho also link to the RoMEO sites which are links for referral regarding questions on copyright.[47]

Communication plans organise and focus the planning and development of advocacy activities. The existence of communication plans were mentioned by two cases, Minho and Connecting Africa, which reflects the significance of and commitment to advocacy by repository management. Minho's plan was the first activity in its advocacy programme which formed the basis of all of its advocacy activities.

Celebrating achievements

A means to advocate ambitions, successes and plans to both the open access and research communities is by officially launching services or repositories or by celebrating the achievement of milestones in the life of a repository. This was reported by several cases interviewed. Cream used the esteemed CNI (Coalition for Networked Information) Conference as a backdrop to launch its service in front of a prestigious open access audience with many of its contributing authors in attendance. CA on the other hand launched its service as part of a conference it organised entitled 'Bridging the North-South Divide in Scholarly Communication on Africa. Threats and Opportunities in the Digital Era', where the researchers and CA staff entered into a dialogue on scholarly communication and open access using the CA repository as a case in point. The success of this event has resulted in plans to make it one in a continuing series. Minho University actively celebrates repository milestones on a regular basis. Minho's institutional repository and its establishment one year after conception was marked in the grand hall of the university with its rector and many faculties present. Now, on an annual basis, Minho continues to celebrate this date and has hosted two conferences with leading open access advocates from abroad. These events have served to share good practices and discuss key open access issues with the Portuguese library and research communities. Such events can stimulate the population of repositories by raising awareness of ambitious and successful institutional repository stories which can encourage contributor participation.

Advocacy as an adhesive for project goals

Advocacy is important on various levels, to gain high-level support for the concept of the institutional repository and its development, and to support the individual in deposit or to persuade late adopters. However, it is also important to improve on or further solidify organisational networks for the successful achievement of project or service goals. This was seen with Cream where efforts were joined for efficiency and team building to further ensure targets were met. The Dutch institutional repository DARE community had a communication network of communication experts from Cream

partner institutions who considered communication issues to both support the institution in advocating the project locally as well as nationally or internationally. A leaflet was designed and a letter drafted to inform authors of the project's aims, and role-playing sessions were staged to prepare the DARE/Cream community for challenges with authors. In addition, communication between a number of institutions was promoted by SURF by setting up an internal DARE community website where project targets, events and documents were shared. A project newsletter updated on a monthly basis was circulated by email to all DARE Community participants involved informing them of progress made and highlighting achievements and reminding them of the goals ahead.

The extent of advocacy and the influence on populating repositories

The extent to which the cases studied have used advocacy differs as seen above, and the importance they place on it is reflected. 'Advocacy has been absolutely key to our repository – without it there would be no full text in the repository – or a small amount from keen people who wished to deposit anyhow,' says White from Soton. Minho also sees advocacy activities as having been successful in contributing to institutional repository population. Little advocacy has been used in the case of the institutional repository at ECS however, with a (computer science) community convinced of the importance of open access. Les Carr as repository manager sees it more as a responsibility and a given to comply to the School's mandate to deposit now. However, more advocacy, at least surrounding the provision of new services, may be necessary in order to achieve the 100% goal, which is a thought shared by CERN who believes that a mandate alone will not achieve this aim. CERN confesses to spending very little time on active advocacy within CERN and more on retrieving content from external sources and open access publishing and now hopes to step up advocacy by presenting it more structurally at departmental meetings. SURF felt that their communication initiatives were highly successful, and it was the product itself as well as how it was sold which helped considerably in bringing real attention to repositories in a new way: 'The Cream brand gave the repository a face', said the community manager for SURF and Cream Annemiek van der Kuil. To summarise, the difference between how the repositories and services studied have approached advocacy relates to their scopes and corresponding challenges.

Recommendations

The following pointers to stimulate the population of repositories have been identified which relate to advocacy and communication issues. For the complete list of guidelines see section 3.5.

1. Target advocacy activities to
   a. senior management to obtain high-level support. Use them to develop policy or services in order to fill gaps in repository stocks. Encourage them to establish mandates to deposit full text where possible, and couple these with financial incentives to deposit where possible.

If mandates are not yet viable, then implement incentives such as services to win hearts and minds. In the case of institutional repositories, strive for the repository to become a research management information tool or CRIS giving the institutional repository a dual function. One-time registration of the recent academic record for internal evaluation can be combined with the repository function as store and dissemination platform. Lastly, use senior management as a sounding board to further develop your repository and its policies in the future.

   b. all other stakeholders involved in the deployment process, be they research heads, researchers, research information administrators, secretaries, etc. Use arguments such as depositing in the repository or CRIS is obligatory (if relevant); deposit will increase worldwide visibility in general, as well as visibility in search services of scientific relevance and thereby increase research impact; this contributes to the open access movement – the new advance in scholarly communication; service X will solve concrete researcher problem Y.
   c. support a project or service network in realising ambitious goals

2. Be clear about what open access stands for and the benefits that the repository has for the researcher. Be informed about the history and practice of open access, at best quoting examples of those who deposit with whom researchers can identify with, for example, colleagues, well-known figures and other institutions. Ensure that the repository and its services address real researcher needs or problems, and review these issues at regular intervals. Be clear as to the relation between your repository or service to other repositories or current research information systems and aim to make links to these for maximum efficiency.
3. Showcase your efforts and achievements by marketing your research results by publicising recent repository additions by research area/department feeding them back to authors and/or research department. Install log analysers of download and upload statistics and pass on this data to depositors and their research groups. Promote the dissemination of this information via departmental websites, individual web pages, etc.
4. Celebrate milestone moments in the development of your repository by organising expert meetings, discussion fora, sharing your progress and challenges with the research and information professional community.
5. Develop a communication plan to identify 1) your target groups, 2) challenges in communicating with them, and 3) specify communication tools to resolve those issues within a set time frame.
6. Consider the best means of acquiring missing content before investing in advocacy efforts.

3.4.6 Legal issues

There is no question that intellectual property rights have a significant effect on the population of repositories. Authors clearly do not want to jeopar-

dise relations with those who contribute to the evaluation of their work. It is clear that there is still a lack of awareness in the research community as to the options available in the changing world of open access scholarly communication. Through education, being aware of the issues at stake can bring about action and cultural change so that repository managers and authors alike can better ensure deposit of current research results now and in the future. This section will focus on present IPR obstacles to populating repositories and work-arounds as well as stimuli for depositing material.

Hampering the population of digital repositories

Lack of awareness

Minho, Soton and HAL all pointed out that author inertia is an inhibiting factor to populating repositories. This is partly caused by the lack of knowledge surrounding author rights and opportunities related to self-archiving and copyright. Fears also exist that copyright may be abused when depositing material in a digital repository. Authors clearly do not want to jeopardise current relations with publishers and awareness-raising efforts are therefore vital. CERN, for example, indicates that post-prints are not always posted by authors as they are unaware as to whether publishers allow such deposits. HAL pointed out that it is also the information and library professionals who sometimes lack the relevant knowledge to inform. For this purpose, Cream of Science had a legal advisor inform the Dutch network of institutional repositories. This was of great value for advocacy efforts and stimulated reluctant authors to collaborate when informed with authoritative answers to legal questions. Institutional repository staff has the responsibility therefore a) to be well informed and b) to pass that knowledge on to its authors. All cases studied are carrying out aactivities in this area.

Know your rights

To guarantee deposits in the future it is the repository manager's responsibility to educate the author in his/her publishing rights and opportunities. For this reason, researchers need to be made aware of the choices of publication and access rights before signing all rights away to the publisher. Some libraries are being proactive by informing the author of his/her rights to self-archiving when negotiating with publishers. Minho does this by sending out informative emails to its authors as a means to better secure self-archived copies of academic output in the future. In addition to this, all cases interviewed are making efforts to inform researchers of the opportunities open to them and to raise awareness accordingly. They do so through their advocacy programmes by providing deposit guides, workshops, FAQs, advice and the like. The importance of legal concerns can be seen at Minho. Despite its informative documentation and support services, 90% of the questions posed to its helpdesk relate to legal issues.

Self-archiving is not permitted or embargoes limit access

Although many publishers allow the self-archiving of material, it is also a fact that self-archiving is not permitted by some, or embargoes exist of a period of six months to two or even three years in some cases after publication. Evidence has shown in this study that material is not deposited in institutional repositories due to publisher policies which do not support access to post-prints or publisher PDFs, for example. CERN and CA point out that population efforts can be slowed by the lack of access to the researcher's full-text version of choice or even to the one permitted by the publisher which is no longer available to the author e.g. post-print for example. This can then have the effect, complains Connecting Africa, of leading readers to mere metadata and not to the free, open access full texts. This has had negative consequences for institutional repositories in providing openly accessible full-text content, and is preventing libraries from acquiring 100% coverage of their academic output.

Self-archiving policies are unknown

All repository managers agree that awareness is crucial to prevent reluctance to deposit. SHERPA/RoMEO claims to contain the self-archiving policies of over 90% of all journal publishers.[48] Minho has therefore implemented a service which allows authors to consult the self-archiving policies of the SHERPA/RoMEO database via the Minho repository web interface before they deposit work. However, self-archiving policies are not always known for certain specialised fields such as African studies, linguistics, humanities, and the like, in addition to journals of non-English speaking languages. Here, in the case of Minho, authors can submit an online form requesting self-archiving policy information on Portuguese journals which are not available as part of SHERPA/RoMEO, which the library will then investigate. Efforts have been made by services such as Connecting Africa to uncover such missing policies on behalf of its users. Evidence has shown that some publishers are not collaborating by not responding to such research efforts, so perseverance is necessary. Some libraries are also reaching out to publishers on behalf of their authors (as part of the Connecting Africa project), requesting the deployment of individual works as an alternative. Although publishers as a whole do respond positively to the self-archival of individual works, these efforts are time consuming and not scalable. Minho on the other hand utilises the letter generated by the COMA project in the Netherlands to semi-automatically request publishers' permission for the deposit of individual articles by its authors.[49] This functionality was adapted and utilised by the library in response to author requests and is another means to aggregate permission to self-archive material. Alternatively, for the future stability of repositories and open access scholarly publishing, pressure needs to be put on the publishing community by the authors themselves. This can only be done if the authors are informed of the opportunities open to them as mentioned above.

Authors who publish in monographs or chapters have IPR concerns of their own, which affects population figures of this type. Copyright on books is also more complex with financial implications in providing open access copies. This can have negative consequences on the provision of access to material in certain domains such as the humanities for example which frequently publishes such works. This was expressed as a concern by Southampton. However, publishing houses can, says White of Southampton, be requested on a case by case basis to make such works, or parts of them, open access. This method has seen more successes than the policy investigation activities to date for many, and should therefore be considered if coverage and population has priority, and is probably the order of preference for the medium of books.

Securing agreements

Another means to vanquish fears surrounding legal issues can be by securing legal agreements between author and library on the material to be deposited. Some cases interviewed make efforts to secure agreements between author and library to deposit material and publish it open access, whereas other do not. This is part of the submission process as part of the DSpace interface at Minho. Here, authors give the library the exclusive right to publish works deposited open access. HAL similarly requests the author to indicate that he/she has the right to post material before doing so. Soton promotes copyright transfer agreements to individual academics although this is not part of the deposit system. Such methods can strengthen trust between authors and their repository managers, which can consequently result in increased deposit.

Abiding by the rules versus serving the researcher

Population targets are clearly affected by the organisation's policy. CERN will only knowingly store versions which publishers allow. This is common practice in many institutions. Soton also has a take-off policy for material which violates copyright agreements. However, in the case of Cream of Science for example, two camps of authors and institutions were identified: the first were those who supported the principles of open access and took a more liberal approach. It was the further visibility of research that had priority as opposed to the commercial interests of the publishers which was sometimes considered to hamper the aims of the dissemination of research. This meant making research openly available regardless of publisher self-archiving policies. This is a means of gaining population targets though it is not a popular one. The second camp was far more cautious and apprehensive of possibly damaging publisher relations. This group was then unwilling to consider depositing anything which went against publisher policies. These differing policies resulted in some authors who had all of their publications online via open access and others with fewer online publications.

Stimulating the population of digital repositories

Local copyright law clearly has an influence on the population of a repository. In the case of the Netherlands, its difference to various other European policies has had a positive effect on the population of its repositories. It was the interpretation of Dutch copyright law related to journal articles produced prior to the so-called 'digitisation era' which has had a considerable effect on institutional repository take-up by the research community in the Netherlands. This interpretation permits the digitisation, storage and dissemination of articles produced prior to 1998 without the need to further consult publishing houses. The Waaijers interpretation was advocated by SURF, leaders of Cream, and its lawyer Wilma Mossink and is still an important stimulus for the development of repository content stocks across the Netherlands. This advice both enhanced author take-up by dispelling author fears of abusing publisher relations on at least part of the researcher's academic output. As a result, a surge of material was digitised for Cream and Connecting Africa services for example. This saw the delivery of personal old hard copies of journal articles for retro-digitisation by the research community resulting in the production and provision of approximately 20,000 new online documents for Cream. Copyright law clearly has a significant impact on the type and amount of content one may be able to acquire for a repository and will therefore colour an institutional repository's collection profile. In the case of the Netherlands, aspects of copyright law have stimulated institutional repository population, or at least have not hindered it; in other countries it has had the opposite effect.

Recommendations

The following pointers to stimulate the population of repositories have been identified which relate to legal issues. For the complete list, see section 3.5.

1. Provide intellectual property rights support by analysing the publisher challenges within your specific subject communities. Ensure that your institutional repository team liaising with the author is informed and up-to-date on self-archiving and related publisher policies. Utilise and monitor tools such as SHERPA/RoMEO to support you in your information
2. Communicate about this issue by admitting to the challenges and fears surrounding IPR; empathise with the author. Point out what can be done rather than what not.
3. Encourage your authors to liaise with publishers on the self-archival of their own work, striving for the immediate deposit of publications in repositories in the future.
4. Liaise with publishers on a case by case basis if time and resources allow.
5. Discuss with your authors how to improve the dissemination of their work, experimenting with them on making more material openly accessible via Creative Commons licences or deposit licences, for example.
6. Secure agreements between library and author where possible.

3.5 Seventeen pointers for stimulating the population of repositories

These pointers are based on the critical success factors and inhibiting factors mentioned in the interviews done for each of the six cases, combined with conclusions made from the research.

1. **Know your research community,** and address their disciplinary needs specifically
   a. speak to the author as part of a particular disciplinary group
   b. speak to young and more mature authors about their differing work processes, challenges and motivations to publish
   c. the mission of the repository and its services should be a tool for the researcher fist and foremost, answering real needs and resolving author and reader problems
2. **Target advocacy activities**
   a. **to senior management to**
      i. obtain high-level support
      ii. implement mandates to deposit full text where possible referring to those organisations already with mandates in place
      iii. implement financial incentives to deposit where possible, referring to those organisations already with them in place
      iv. if mandates aren't possible immediately, then implement incentives for deposit such as services to win hearts and minds
      v. develop policy or services in order to fill gaps in repository stocks
      vi. strive to become a research management information tool or CRIS, giving the repository a dual function, i.e. one-time registration of the recent academic record for internal evaluation and storing and disseminating that content via the repository
      vii. use them as a sounding board for policy and repository development in the future
      viii. if national funding bodies mandate deposit, use this argument when presenting the case for deposit.
   b. **to all other stakeholders involved in the deployment process**, be they research heads, researchers, research information administrators, secretaries, etc.
      i. use arguments such as
         – using the institutional repository increases worldwide visibility through online open access and more visibility in generic and disciplinary search engines and services
         – increase the impact of research
         – service X will solve researcher problem Y (see below)
         – contribute to the Open Access movement
         – depositing in the repository or CRIS is obligatory (if relevant)
   c. **to project or service contributors** to support the network in realising ambitious goals

   d. **and develop a communication plan** to identify your target groups and the challenges in communicating with them, and specify communication tools to resolve those issues within a set time frame
   e. **while carefully considering the best means of acquiring missing content** before investing in advocacy. Sometimes advocacy efforts can be limited if other means are used to acquire missing material from outside
3. **Be clear about what open access stands for and what benefits the repository offers to the depositor**:
   a. clearly inform authors on the history and practice of Open Access in various communities; cite examples researchers can identify with, e. g. colleagues, well-known figures, competitive institutions and others who deposit
   b. ensure that your repository and its services address researchers' real needs or problems, and review these issues at regular intervals
   c. be clear as to the relation between your repository or service and others utilised by your researchers, e.g. a disciplinary repository or CRIS, and aim to make links to these for maximum efficiency
4. **Make collection development choices which reflect the academic output of your disciplines.** Home in on the challenges in unlocking that material, as self-archiving traditions and possibilities differ widely across disciplines.
5. **Provide added value services which are flexible and adaptable to save the researcher time on non-research activities**
   a. strive to ensure that very little time and effort is needed to deposit material
   b. demonstrate repository services or systems in the ideal case
   c. provide feedback on the use of the deposited material to the depositor
   d. examples of services are:
      – CVs or automated publication lists
      – search and browse facilities to allow as much cross-interrogation as possible, with the possibility to discover not only new research, but networks, be they institutional or people
      – push out information from your repository to disciplinary services on behalf of your researchers
      – use RSS feeds to feed back new material that enters the system to various disciplinary groups
      – retro-digitise older material
      – seek to preserve the academic record, convert formats, and seek to implement long-term programmes in collaboration with others
   e. monitor the cost-effectiveness of these services
6. **Take on an active role in improving on information retrieval and discovery by**
   a. contributing your content to services with a regional or international significance and feeding back results to your researchers
   b. targeting information services of significance such as Google, Google Scholar, as well as disciplinary ones

c. aiming to optimise the positioning of your material in the result sets of these information services by taking this up with the service providers

7. **Push out your content to the world research community to show your commitment to increasing the impact of your researchers' work**
   a. liaise with your researchers to identify which sources are of significance
   b. make efforts to get your repository indexed or harvested by these information services
   c. try to optimise the positioning of your material in these sources
8. **Showcase your efforts and achievements**
   a. market your latest repository research result entries by publicising recent additions and usage statistics feeding them back to authors and/or research departments
   b. highlight the most recent additions to your repository by research area and/or department
   c. provide results to departments, and encourage dissemination via departmental websites, individual web pages, etc.
   d. install log analysers of download and upload statistics to help market your repository
   e. celebrate milestone moments in the development of your repository by organising expert meetings and discussion forums to share your progress and challenges with the academic and information professional community
9. **Be innovative as to how you acquire your content. If at first you don't succeed, try, try again**
   If your researchers deposit elsewhere:
   a. identify which archives are places of deposit for the researcher of a particular discipline
   b. make agreements with selected sources to either harvest metadata or full text depending on the mission of your repository
   c. monitor that data
10. **Provide intellectual property rights (IPR) support by**
   a. admitting to the challenges and fears surrounding IPR; empathise with the author
   b. emphasising what can be done rather than what not
   c. analysing the publisher challenges within your specific subject communities where different challenges will be apparent
   d. ensuring that your institutional repository team liaising with the author is informed and up-to-date on self-archiving and related publisher policies
   e. utilising and monitoring tools such as SHERPA/RoMEO to support you in your information
   f. liaising with publishers on a case-by-case basis if time and resources allow

g. encouraging your authors to liaise with publishers on the self-archival of their own work, striving for the immediate deposit of publications in repositories in the future
h. discussing with your authors how to improve the dissemination of their work in the future and experimenting with them on making more material open access
i. securing agreements between library and author where possible

11. **When organising your repository**
    a. consider the distributed organisation of academic output within the institution when planning population
    b. consider the research organisational structure of your organisation and adapt to it
    c. strive to give some autonomy to the research community by giving them the responsibility to maintain their output with library support. Encourage them to feel like the owners of the institutional repository output, and provide them with an interface that matches the look and feel of the department(s)
12. **Ensure that the infrastructure is in place to deliver**
    a. ensure that the staff is available to provide the necessary support
    b. ensure that services can be developed to support the researcher
13. **Use your local, regional, national and international networks for**
    a. policy development
    b. service development
    c. personnel development
    d. publicising your work
14. **The repository team needs to**
    a. be a strong, knowledgeable team
    b. provide a simple deposit system
    c. have sound knowledge on intellectual property rights and self-archiving practices
15. **If you seek to develop a regional or disciplinary service to help populate your repository**
    a. choose a prominent partner with influence, preferably well recognised by the library or research community to lead it
    b. aim to provide support on an operational level, as co-ordinator
16. **Strive for cost-effectiveness by**
    a. analysing work processes and striving for synergies with the CRIS or research management information department or with funding agencies who mandate repository deposit
    b. striving for departmental/researcher deposit and not library deposit for long-term sustainability; however, do consider repository investment in acquiring content and investing time to provide a good demonstrator to encourage self-deposit in the future
    c. share experiences with colleagues in a similar position, be they repository managers, policy makers, technical developers or communicators

d. harvest content from outside your repository to further acquire missing content
e. use SOAP[50] services to update content

17. **Challenge yourself**
    a. just do it and don't spend too long thinking about it
    b. be willing to risk a new idea

3.6 Conclusions

Conclusions are based on six digital repositories and services which are different in nature, scope and model. Despite this very small sample, a number of lessons can be learnt on how to populate a digital repository from these European good practices, which are relevant to many worldwide. Some conclusions do echo trends in the repository world, and it is some of these cases studies who have helped set them. The institutional repository community will decide how to utilise knowledge gained from these cases for their own local situations. The cases' geographical, cultural and organisational contexts in addition to the subject communities the cases serve need to be considered here. Repository managers and policy makers will profit from reading the in-depth case studies for a detailed analysis of particular repository models which are close to their own experiences. These can be found at www.driver-repository.eu.

This study has identified seventeen recommendations which have been taken from the good practices. They can be considered as examples of stimuli for improving on the population of repositories and their services. They fall under the following six areas of study.

As regards **policy issues**, evidence in this study supports Harnad's and Sale's push for institutions to mandate the electronic deposit of academic output. Numbers have indeed increased on the implementation of such mandates as seen at Minho, Southampton and CERN. Mandates should therefore in an ideal case be striven for. However, those organisations with mandates in place also emphasise that to reach their 100% academic output ambitions, the development of incentives will also be necessary through service development, for example. This study also shows evidence that service development for a specific research community, be it of a scientific or national character, can also enhance the population of a repository as seen with HAL, Cream of Science or Connecting Africa – all places where mandates play no role, and authors are encouraged via incentives. Both mandates and service incentives therefore need equal consideration.

Knowledge exchange and networking is also significant to the cases studied. Local, regional, national and international networks have served to develop and strengthen policy, been an inspiration for developing services, and has stimulated personnel development and adhesion in projects. For example, SURF, as leaders of Cream and the national DARE institutional repository network, saw its separate networks of repository managers, pol-

icy makers, technical and communication experts sharing expertise on IPR issues, technology and advocacy as a critical success factor for the population of its institutional repositories nationwide. Such a network model, led by an influential body, can have great effects on achieving high aims in a cost-effective manner.

As far as **organisation** and its influence on repository population is concerned, it is vital to ensure that the infrastructure, that is, staff and services, is in place to deliver results. What is clear is that it is important to have a repository with a simple deposit system and a knowledgeable support team, particularly in the area of IPR. When organising a repository and its population, it is above all vital to thoroughly consider the way that the documentation and deposit of research publications is organised and the role of the research department or faculty within that process. This may be on an institutional level when talking of institutional repositories, or on a broader level when more than one repository is involved in feeding a service. Analysing work processes will enhance engagement by the research community as well as increase efficiency. Reflecting this, repositories with a dual function as a current research information system (CRIS) – linking the obligatory documentation of research results with full-text deposit – can further guarantee institutional and researcher support. This can then result in the increase of the repository's or service's current bibliographic references and full text.

Concerning the drivers necessary to lead and cultivate repositories, there is no doubt that involving high-level management in policy and service establishment and development will create institutional buy-in and will thereby encourage population numbers to increase. However, if the central organisation of the academic output on an institutional level does not succeed, other models can reap results. For example, giving autonomy to the research community in defining repository policy and in being responsible for the further dissemination of their work through that can create a new feeling of responsibility within the research community for making a positive change to the scholarly communication process as can be seen at Minho. On another organisational level, if a regional or disciplinary service is set up, which in turn helps populate repositories, a partner of influence needs to drive the initiative. This authority will see more cooperation from deployers.

There are various **mechanisms and influential factors for populating repositories**. The repository manager should endeavour for cost-effectiveness when populating a repository now and in the future, which will further secure institutional support. This can be done by analysing work processes considering the documentation of academic output, linking in to other archives, harvesting content from others, and by using web services. Researchers are almost certainly more willing to deposit were repository efforts to clearly reflect their communities and needs. For this reason, also being aware of the differences in self-archiving traditions and possibilities and in the challenges of unlocking that material across disciplines will

further support efforts to motivate deposit. Making repository collection development choices should similarly reflect the academic profile of the existent disciplines. Challenges clearly still exist in obtaining critical mass, and it is important to be innovative when acquiring content. If there are delays, organisational issues need to be dealt with or services need to be developed to acquire that content. However, it is also important to endeavour to find other ways of reclaiming content. For example, through harvesting or by looking at IPR opportunities focussing on publishers which allow the open access deposit of publisher version material.

Once researcher needs and issues have been identified, it is through the development of **services** built upon repository content which is where the archive can bring added value to the researcher's research life. It is therefore advised to provide services which save the researcher real time on non-research activities such as one-time deposit for multiple dissemination, i.e. via departmental or individual web pages, automated publication lists, search engines and disciplinary search services. A prerequisite for this is a simple and rapid deposit system. The repository manager should take on an active role in improving on information retrieval and discovery of its author's work by ensuring that repository material appears in renowned search engines such as Google or Google Scholar and he/she should aim to optimise the position of the author's work within them. In addition, investigating the information sources in which the author hopes to be located on a disciplinary level is also vital. This thereby shows the commitment to increase the impact of the researcher's work. Aside from this, if preservation is a selling point, ensure that an infrastructure and action plan exists and is in motion to be able to fulfil expectations. Evidence shows that authors are willing to contribute their content partly due to service developments – be they portals based on a number of repositories such as Cream of Science – or sub-services of repositories such as publication lists. The involvement of more users in the design of services as seen in Connecting Africa, and the analysis of their take-up, and a concluding cost-benefit analysis could well increase the efficiency of institutional repository investments.

Marketing repository or service efforts to all stakeholders involved in the deployment process, be they research heads, researchers, research information administrators or secretaries, will help achieve ambitious results. Clear arguments are needed which answer real problems. Above all be clear as to what Open Access stands for and the direct personal benefits that the repository has for the institution, the author and the potential future reader of that material. **Advocacy and communication** activities should ideally be defined in a communication plan, which will help identify target groups, challenges in communicating with them, and specify communication tools to resolve the issues of population.

For institutional endorsement when setting up a repository or service or when introducing important new policy, target advocacy activities to senior management who can make institutional cultural changes. Utilise them to establish mandates and incentives to deposit academic output open access, also preferably aiming for the repository to take on a dual role as a CRIS. It is clearly essential to understand the research community and address them both differently and specifically. They need to be addressed understanding a) the specificity of the discipline where self-archiving traditions and concerns can vary and b) the motivation for publication and importance of it in the researcher's career path to reap the best results. Empathising with the researcher's needs and challenges and seeing how the repository can directly resolve concrete issues are imperative for researcher take-up. Showcasing achievements by publicising faculty publication results, usage statistics which show the use of material deposited, as well as milestone successes in the development of the archive or service will motivate. Separate advocacy programmes can also be used in order to inform, motivate and achieve ambitious goals amongst repositories for the realisation of projects or services. However, it is advised to weigh the significance of advocacy programmes in acquiring content if other methods prove to be more viable such as harvesting or reclaiming material from outside the institution to increase population figures.

Legal issues clearly hinder the deposit of academic output at present, and providing intellectual property rights support is therefore essential. Repository managers should take a more active role in raising awareness of the self-archiving opportunities available such as immediate repository publication deposit or the authorised self-archival of specific versions by large publishers. Information must be up-to-date and accurate. This needs to be done with sound knowledge, and ideally with the support of a legal authority. Further trust can thereby be gained by the research community in this complex area. A balance needs to be found between preventing copyright breaches and challenging the present scholarly communication mechanisms in more widely disseminating research output open access.

In summary

Current incentives and mandates are not reaching the 100% mark. Mandates will undoubtedly help us populate our institutional repositories; evidence has shown this. However, winning the hearts and minds of the researchers through services which are modelled on supporting the researcher in his/her work processes will build on and forge new relationships between the research community and the information professional one. This is vital in sustaining and confirming our non-commercial role in providing support to the needs of research in the rapidly changing digital age of scholarly communication. The specificity of the discipline should not be underestimated in the population of our repositories as affinities to the goals of self-archiving and the abilities to comply vary diversely. This means

that it is important that we target the research department rather than the merely the faculty or institution.

Population challenges would certainly dwindle were repository managers to take on the responsibility for disseminating the academic output of their authors in a broader capacity than has been shown to date. Services based on repository content can guarantee the structural dissemination of research results via various media of importance to the researcher. One-time repository input should result in access to that data via the researcher's web page, departmental and university website, management summaries such as annual reports, in research evaluation outputs, subject-specific online services, generic online search engines, and so on. Evidence has shown that repository managers are developing some of these services; however, giving the repository a new function as being *the* formalised institutional source and responsible body for the further visibility of research output could truly root the repository in the new world of scholarly communication. This is an ambitious challenge.

Evidence is now here on some of the strategies which will enhance institutional repository population, this report expands on a mere six. It is our responsibility to be alert to these and others and to select the lessons which will help us on our journey to gain critical mass across all disciplines as we have not yet reached our destination.

4. Intellectual property rights

Wilma Mossink

4.1 Introduction

Within the DRIVER project, work package seven consists of several introductory studies regarding topics relating to setting up and maintaining digital repositories. Difficulties with solving copyrights problems often hamper the filling of digital repositories and hinder a smooth management of the repository. Questions about ownership of the works in the repository, or getting permission to use a work in accordance with the ideas and principles of open access frequently take a lot of time or even hinder the creation of a fully accessible repository. Examples or models to overcome the copyright problems would be most helpful. Therefore, within the context of the DRIVER study, a study on intellectual property which provides examples and models is indispensable.

The aim of this inventory study is threefold. Firstly it gives an overview of copyright and other intellectual rights relevant for digital repositories. Secondly it provides insight in what (study and experimental) work on intellectual property is already being done in the EU. Furthermore the study provides models to continue working with and for developing digital repositories in the EU context, in order to arrive at sustainable development and operation at the local, national and international levels. It is written for those who are involved in the process of setting up a repository, or those who are running one.

The starting point of this study is the central position of the author in the landscape of scholarly information. It examines the legal relationships an author has to go into to make his work fully openly available. The background for this study is the open access principle as articulated in various declarations regarding open access. Three different relationships are the subject of the study: the legal relationship between the author and his/her institution, the legal relationship between the author and the publisher, and the legal relationship between the author and society.

The relationship author-institution will cover the topics of copyright ownership of scholarly publications and copyright policies. The strand author-publisher goes into the matter of the key needs for authors and publishers, publishing agreements and the principles an author might take into account when publishing an article in a journal that is not an open access journal. The relationship author-society is a more complex one. The terms and conditions under which a work can be made available will be discussed. The digital deposit licence is the most important topic here.

Appendix 3 comprises an overview the relevant European initiatives and good practices. In this appendix, known names and contact details of persons in the European countries who are engaged in the legal aspects of digital repositories are mentioned.

4.2 Intellectual property rights explained

4.2.1 Copyright

This part of the study describes briefly those elements of copyright and database rights a repository manager should take into account setting up/managing a repository. After reading this part the repository manager will be informed as to which and when a work is protected by copyright. He will know how copyright of a work can be exploited and how the database right could affect the repository. He will comprehend the concept of exclusivity of a licence. Understanding copyright and realising how to exploit those rights can help one to manage a sustainable digital repository. Because of the different law systems known in Europe the differences that stem from the legal systems of these countries regarding the exclusive rights of an author and the way these rights can be exploited must be handled differently. Therefore, for each topic discussed in this legislative overview, four countries will pass under separate review. The situation in the Netherlands will be discussed as a starting point. Germany and France will follow as typical exponents of civil law countries. Finally, the United Kingdom will be discussed as example of a common law country.

Copyright
Copyright is the exclusive right of the author to reproduce and distribute his work. Copyright in relation to a work means that the owner of the copyright in a work has the exclusive right to perform certain actions. Those actions are enshrined in a copyright act or copyright law. The exclusive right can be an economic or a moral right. Basically economic rights give access to a work, whilst moral rights protect the tie between the maker and his work.

To get copyright an author does not need to register his work. The enjoyment and the exercise of the rights of an author are not be subject to any formality.[1] Copyright exists from the moment a work has been created. The term of copyright protection ends 70 years after the death of the author. After that term, the work enters the public domain. The author or first owner of a work's copyright is the person who created the work. Exceptions to this principle are found in some countries when a work is made for hire or a work is made in the course of employment. In that case the employer is the first owner of copyright. This principle initiated many debates about the ownership of scholarly works. This will be further discussed in the part of this study where the relationship author and institution is discussed.

A work has more owners than just one in the case of joint ownership. Apart from the Netherlands where the criterion for joint ownership results from jurisprudence, the copyrights act of Germany, France and the UK have provisions for this which all bear resemblance to each other.[2] The Dutch Supreme Court has ruled that joint ownership exists in cases where the different elements in a work cannot be divided and when the elements cannot be judged separately. This rule resembles §8 of the German Act where two or more authors jointly produce a work in which the contribution of each author cannot be distinguished from those of the other authors, their work is a work of joint ownership. The legal consequence of joint ownership is that the authors can only exercise the copyright jointly. Nevertheless each author is entitled to assert claims arising from infringements of the joint copyright.

Economic rights

The Dutch Copyright Act describes two exploitation rights: the right of communication to the public and the reproduction right. The act gives no explanation of the communication right; article 12 of the act merely sums up what the right of communication to the public entails.[3] By communicating to the public an author makes his work accessible so that the public can take notice of it. The act does not explain the reproduction right either but the reproduction right is seen as the act of producing copies of the work.[4] Basically two forms of the reproduction right can be distinguished: creating one or more physical copies of the work, and making an adaptation.

Germany knows one right in which the economic and moral rights are intertwined.[5] Apart from the moral rights, the author enjoys four categories of rights. These are rights of exploitation, rights connected with use rights, other rights and additional rights of remuneration.

The economic rights belonging to the author in France are comprised of the right of performance and the right of reproduction. According to the Intellectual Property Code, the right of performance consists of the communication of the work to the public by any process whatsoever. The Code lists public recitation, lyrical performance, dramatic performance, public presentation, public projection and transmission in a public place of a telediffused work and telediffusion.[6] Also transmission of a work towards a satellite is considered a performance.

The reproduction right consists of the physical fixation of a work by any process permitting it to be communicated to the public in an indirect way. It may be carried out, in particular, by printing, drawing, engraving, photography, casting and all processes of the graphical and plastic arts, mechanical, cinematographic or magnetic recording.

The adaptation right and the right to control the use and circulation of copies are incorporated in the right of reproduction. Sections 16-21 of the English CDPA 1988 set out economic rights that are currently granted to copyright owners. The copyright owner has the exclusive right to do the following acts in the United Kingdom: copying the work (reproduction right), issuing copies of the work (distribution right), renting or lending

the work (rental or lending right), or adapting the work (right of adaptation).

Moral rights

Moral rights protect the bond between the maker and the work he created. This bond is so personal that it needs to be safeguarded. Moral rights are a concept of the civil law system; in common law moral rights are more limited than in civil law jurisdictions.

The moral rights enshrined in the Dutch Copyright Act give an author the right to oppose communication to the public of his work without acknowledgement of his name or other indication as author, or under a name other than his own, in so far as it appears on or in the work or has been communicated to the public in connection with the work. Furthermore, the author has the right to oppose any alteration of his work as well as any distortion, mutilation or other impairment of the work that could be prejudicial to his name or reputation or to his dignity as such. These rights are maintained even after the author has assigned his copyright.

In the United Kingdom, the concept of moral rights only slowly precipitated into legislation because of the applicable EC Directives. It was not until 1988 that the United Kingdom copyright legislation gave specific recognition to moral rights. The CDPA 1988 provides the right to be named when a work is copied or communicated (attribution right), the right to control the form of the work (integrity right) and the right not to be named as the author of a work which one did not create (right to object to false attribution).[7] The right of attribution is granted only to creators of original literary, dramatic, musical and artistic work and films and does not apply to software. An author must assert his right to attribution before it can arise. This can be done in two ways: to include a statement that the work has been asserted or by a written document signed by the author.

An author in Germany has three different moral rights. The first is the right to decide whether and how his work is to be published (right of publication). An author has the exclusive right to publicly communicate or describe the content of his work, as long as neither the work nor its essence nor a description of the work has been published with his consent. The second constituent of the German moral right is the recognition of authorship. The author may decide whether the work is to bear his designation and what designation is to be used. The third component of the moral right is the integrity right. An author has the right to prohibit any distortion or any other mutilation of his work which would jeopardise his legitimate intellectual or personal interests in the work.

A French author enjoys the right to respect for his name, his authorship and his work. This right is attached to his person and is perpetual, inalienable and imprescriptible. It may be transmitted mortis causa to the heirs of the author. Furthermore the author shall have the right to divulge his work. He shall determine the method of disclosure and shall fix the conditions thereof. Notwithstanding assignment of his right of exploitation, an author enjoys a right to reconsider or withdrawal the work, even after its publica-

tion, with respect to the assignee. However, he may only exercise that right on the condition that he indemnifies the assignee beforehand for any prejudice the reconsideration or withdrawal may cause him. If an author decides to have his work published after having exercised his right of reconsidering or withdrawal, he shall be required to offer his rights of exploitation in the first instance to the assignee he originally chose and under the conditions originally determined.

Protected works

Most of the time one can find which works are protected by copyright in the copyright law. The list of protected works is either exhaustive or non-exhaustive. In the latter case any creation in the literary, scientific or artistic areas, whatever the mode or form of its expression, is protected. Facts, ideas, methods or opinions on which a work is based cannot be protected by copyright; they are in the public domain.

Article 10 of the Dutch Copyright Act provides a non-exhaustive list of works protected by copyright. The act protects literary, scientific or artistic works, which include writings like articles, books, newspapers and periodicals. Not only writings are protected, but also performances, geographical maps, cinematographic, photographic and dramatic works, and so on. Computer programs and their preparatory material fall within the category of protected works, and since the implementation of the Database Directive, databases are also included. Reproductions of a literary, scientific or artistic work in a modified form, such as translations, arrangements of music, cinematographic and other adaptations and collections of different works, are protected as separate works, without prejudice to the copyright in the original work. In the Netherlands no copyright subsists in laws, decrees or ordinances issued by public authorities, or in judicial or administrative decisions.

The German Copyright Act has an exhaustive list of protected works. The act protects literary, scientific and artistic works. These include works of language, such as writings, speeches and computer programs, musical works, works of pantomime, choreographic works, works of fine art, works of architecture and applied art and plans for such works; photographic works, including works produced by processes similar to cinematography; illustrations of a scientific or technical nature, such as drawings, plans, maps, sketches, tables and three-dimensional representations.

Translations and other adaptations of a work enjoy protection as independent works, without prejudice to copyright in the work that has been adapted. Collections of works, data or other independent elements which, by reason of the selection or arrangement of the elements, constitute a personal intellectual creation (collections) enjoy protection as independent works without prejudice to a copyright or neighbouring right existing in the elements included in the collection.

France protects the rights of authors in all works of the mind, whatever their kind, form of expression, merit or purpose. As works of the mind are considered all kind of writings, lectures, sermons and other works of such

nature. Dramatic and musical compositions, films and all kind of graphical works are protected under the Intellectual Property Code. The authors of translations, adaptations, transformations or arrangements of works of the mind also enjoy the protection afforded by the Intellectual Property Code, without prejudice to the rights of the author of the original work. The same applies to the authors of anthologies or collections of miscellaneous works or data, such as databases, which, by reason of the selection or the arrangement of their contents, constitute intellectual creations. In France the title of a work of the mind is protected in the same way as the work itself where it is original in character. Such title may not be used, even if the work is no longer protected to distinguish a work of the same kind if such use is liable to create confusion.

The CDPA 1988 states that copyright is a property right which subsists in original literary, dramatic, musical or artistic works, sound recordings, films, broadcasts, and the typographical arrangement of published editions. The part 'Descriptions of work and related provisions' defines the meaning of the subsequent protected works. A literary work has a very wide meaning and can be written, spoken or sung. A literary work cannot be a dramatic or musical work. In the context of copyright in the typographical arrangement of a published edition, a published edition is the whole or any part of one or more literary, dramatic or musical works.

Originality

In order to be protected a work has to fulfil the criterion of being original. The level of originality differs in the common law and the civil law systems but the level is converging because of the harmonising effect of European legislation.

In the Netherlands a work must be a form of ideas, opinions or feelings of a maker that are perceptible by senses. The work must be original and in addition to this also has to bear a personal mark of the author.

In Germany the threshold for originality is rather low and excludes only very trivial works. Here is the criterion that the work is intellectual and original and that it is a 'personliche geistige Schöpfung' (personal creation of the mind).

The French Intellectual Property Code does not require that the work is fixed in some material form in order to get protection. A lecture that is not written down is protected if the work meets the requirement that it is original. The Intellectual Property Code does not mention this criterion but jurisprudence and doctrine have developed it.[8] An original work has to wear the mark or personality of the author.

In the United Kingdom a work must satisfy several requirements before it is protected by copyright. The work must fall in one of nine categories listed in the CDPA. It also must be recorded in a material form and must be original. The threshold for originality has been set at a very low level. For the purpose of copyright law originality does not mean that a work is inventive, unique or novel; originality simply means that the author must have exercised the requisite labour, skill or effort to produce the work. A difficult

question arises when a scientific work that only reveals facts and data, which are hidden in nature or logic, and which have no personal character, is protected by copyright. But because the reflection of these facts and data in a scholarly work often is the result of a creative process it can be concluded that such a work is protected by copyright.

Exploitation of rights

An author can exploit his economic rights in a variety of ways. He can exercise all the rights himself but he also can choose to let others exploit his work through transfer of his rights or by means of a licence, giving others permission to exploit his rights, exclusively or non-exclusively. A non-exclusive exploitation right entitles the rights holder to use the work, concurrently with the author or any other entitled persons, in the manner permitted to him. An exclusive exploitation right entitles the right holder to use the work, to the exclusion of all other persons, including the author, in the manner permitted to him.

In the Netherlands transfer of copyright is regulated in the second article of the Copyright Act. The delivery required by whole or partial assignment shall be effected by means of a deed of assignment. The assignment shall comprise only such rights as are recorded in the deed or necessarily derive from the nature or purpose of the title. Transfer of rights for works that will come into existence in the future is only possible if the definiteness of the work is defined sufficiently. Wording like 'transfer of rights of works which an author shall make' is regarded as sufficient. Moral rights cannot be transferred, only waived, except for the right to oppose to any mutilation of the work. This right always stays with the author.

Because in Germany the economic and moral rights are intertwined and are not separable, legal consequence is that the economic rights cannot be assigned. Copyright may be transferred in execution of a testamentary disposition or to co-heirs as part of the partition of an estate. Otherwise copyright is not transferable. An author merely grants user rights while the right itself stays with the author. He may grant the right to use the work in a particular manner or in any manner. He can grant this right as exclusive right or non-exclusive. If the types of use to which the exploitation right extends have not been specifically designated when the right was granted, the scope of the exploitation right shall be determined in accordance with the purpose envisaged in making the grant. The grant of exploitation rights for as yet unknown types of use and any obligations in that respect shall have no legal effect.

The French Intellectual Property Code has extensive provisions concerning transfer of rights and publishing contracts. A principle hereby is that the author is entitled to a proportional participation by the author in the revenue from sale or exploitation of the work. To transfer author's rights in France each of the assigned rights must separately be mentioned in the instrument of assignment. Furthermore the field of exploitation of the assigned rights must be defined as to its scope and purpose, and as to place and to duration. Assignment by the author may be total or partial. Any as-

signment clause affording the right to exploit a work in a form that is unforeseeable and not foreseen on the date of the contract shall be explicit and shall stipulate participation correlated to the profits from exploitation. Publishing contracts shall be in writing. The Intellectual Property Code describes a publishing contract as a contract by which the author of a work of the mind assigns the right to manufacture or have manufactured a number of copies of the work under specified conditions to a publisher, in order to ensure publication and dissemination thereof. For this the author pays to the publisher an agreed remuneration. A clause by which the author undertakes to afford a right of preference to a publisher for the publication of his future works of clearly specified kinds is lawful. However this right is limited, for each kind of work, to five new works as from the date of signature of the publishing contract concluded for the first work or to works produced by the author within a period of five years from that same date.

In the United Kingdom copyright is transferable by assignment, by testamentary disposition or by operation of law, as personal or moveable property. The assignment is not effective unless it is in writing signed by or on behalf of the assignor. The assignment or other transmission of copyright may be wholly or partial. A partial assignment is limited to part of what the copyright owner has the exclusive right to do and is limited to part of the period for which the copyright is to subsist. Copyright can also be licensed, exclusively or non-exclusively. Copyright which will or may come into existence with respect to a future work or class of works or on the occurrence of a future event can be assigned or licensed wholly or partially as well. Moral rights are not assignable.

4.2.2 Database rights

Introduction

The exponential growth of the amount of information, in the European Community and worldwide, called for investment in all the member states in advanced information processing systems and the protection of the producers of such advanced systems. The EU feared that the investments in the Community would lag if the producers were not protected, so they initiated a protection right for producers of databases.

As long ago as 1988, the European Commission published a green paper[9] which for the first time drew attention to the legal protection of databases within the context of copyright law and the challenge of new technology. In 1992, the bill for the Database Directive was submitted to the European Parliament. On 11 March 1996 the European Database Directive was introduced.[10]

Database rights are part of this study because repositories are subject to protection under database right. Regardless if the institutions want to exercise this right, they at least need to consider this and take a stand as to how to deal with it.

Database rights

A database is a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means. By way of clarification, the preamble to the Database Directive also states, in Recital 17, that '...the term 'database' should be understood to include literary, artistic, musical or other collections of works or collections of other material such as texts, sound, images, numbers, facts, and data'. This means that a recording or an audio-visual, cinematographic, literary or musical work as such does not fall within the scope of the Directive; these are protected under other copyright laws. Computer programs used in the making or operation of databases accessible by electronic means don't fall under the protection of database right. The Directive applies to both digital and analogue databases; a paper telephone directory is protected just as websites found on the internet.

The Database Directive offers two levels of protection: copyright and a new especially created right called the *database right*. Databases which, by reason of the selection or arrangement of their contents constitute an author's own intellectual creation, are protected by copyright. To attract the database right a maker of a database must show that there has been a substantial investment in either the obtaining, verification or presentation of the contents. That investment may take a variety of different forms. For example, it may consist of the effort, time and energy spent collecting and checking the content of the database. It could also be the work involved in making the database accessible to the public.

The term of protection for a database is fifteen years and runs from the date of completion of the making of the database. Any substantial change to the contents of the database which would result in the database being considered to be a substantial new investment evokes a new term of protection.

In November 2004 the European Court of Justice gave four rulings concerning the interpretation of some of the key elements of the Database Directive: the object and scope of the database protection.[11] The Court gave a narrow interpretation of the object of protection. It decided that an investment in the obtaining of the contents of a database only refers to the resources used to seek out existing independent materials and collect them in the database. The resources used for the creation of materials which make up the contents of a database do not count as a substantial investment and therefore that database would not be protected. The Court also defined the expression 'verification of the contents'. According to the Court 'investment in the verification' refers to the resources used, with a view to ensuring the reliability of the information contained in that database, to monitor the accuracy of the materials collected when the database was created and during its operation. If a repository would substantially invest in the verification of the information in the database, the repository would be protected.

Protection under the Database Directive

The Database Directive provides for the rights of the maker of a database when the maker is a natural person or where the legislation of the member state permits, the legal person designated as the right holder by that legislation. Provided that the maker is a subject of or has his usual place of residence in a member state of the European Union or the European Economic Area, the producer is empowered by law to grant others leave to perform certain actions.

The database right protects the maker of a database against extraction and/or re-utilisation of the whole or a substantial part of the database. Extraction is the transfer to another medium by any means or in any form (reproduction); re-utilisation means any form of making available to the public all or a substantial part of the contents of a database by distribution of copies, renting, online or other forms of transmission. In the aforementioned judgments the European Court of Justice explained to what extent a maker of a database is protected against unauthorised reproduction and re-utilisation. The term substantial part is of significance here. The Court ruled that the terms 'extraction' and 're-utilisation' must be interpreted as referring to any unauthorised act of appropriation and distribution to the public of the whole or a part of the contents of a database. The expression 'substantial part' refers to the volume of data extracted from the database and/or re-utilised and must be assessed in relation to the total volume of the contents of the database. It refers to the scale of the investment in the obtaining, verification or presentation of the contents of the subject of the act of extraction and/or re-utilisation, regardless of whether that subject represents a quantitatively substantial part of the general contents of the protected database. Any part which does not fulfil the definition of a substantial part, evaluated both quantitatively and qualitatively, falls within the definition of an insubstantial part of the contents of a database.

The fact that the contents of a database were made accessible to the public by its maker or with his consent does not affect the right of the maker to prevent acts of extraction and/or re-utilisation of the whole or a substantial part of the contents of a database.

4.3 Landscape of scholarly information

This part of the study shows the different legal relationships an author has in the scholarly environment and how the different relations can be managed to make available his scientific output worldwide. This part can be useful for repository managers when they want to archive and preserve the scientific output of their institutions. It assists the repository manager by giving him some tools that he can use when he tries to convince authors to deposit their works. This part of the study starts with some basic background information about the Berlin Declaration on open access to Knowledge in the Sciences and Humanities, which role digital repositories play in the open access movement, and how repositories contribute to the accessi-

bility of scholarly information. As this study deals with copyright and institutional repositories its focus lies on the legal implications of depositing an article in an institutional repository, which means that open access publishing will not be dealt with.

4.3.1 Berlin Declaration and repositories

The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities defines open access as the worldwide electronic distribution of peer-reviewed literature and completely free and unrestricted access to it by all scientists, scholars, teachers, students and other curious minds.[12] The Berlin Declaration is the third major international statement about digital, online access to primarily peer-reviewed research articles, free of charge, and free of most copyright and licensing restrictions. The first statement, the Budapest Open Access Initiative, arose from a small meeting in Budapest convened by the Open Society Institute. The purpose of this meeting was to accelerate progress in the international effort to make research articles in all academic fields freely available on the internet. The Budapest Open Access Initiative is a statement of principle, strategy and commitment, reflecting on how the separate initiatives in the Open Access movement could work together to achieve broader, deeper and faster success.[13] The Bethesda Statement on Open Access Publishing was released in June 2003. The intention of this statement is to stimulate discussion within the biomedical research community on how to proceed, as rapidly as possible, to the widely held goal of providing open access to the primary scientific literature.[14]

The Berlin Declaration is the one best known, most used and most referred to. This declaration builds upon its two predecessors and is therefore more comprehensive than the Budapest and Bethesda statements. The Berlin Declaration promotes the internet as a functional instrument for a global scientific knowledge base and for human reflection. It specifies measures which research policy makers, research institutions, funding agencies, libraries, archives and museums could consider when disseminating knowledge widely and readily available to society. According to the Berlin Declaration open access contributions must satisfy two conditions:

1. 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 license 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 (community standards will continue to provide the mechanism for enforcement of proper attribution and responsible use of the published work, as they do now), as well as the right to make small numbers of printed copies for their personal use.
2. A complete version of the work and all supplementary materials, including a copy of the permission as stated above, in an appropriate standard

> electronic format is deposited (and thus published) in at least one online repository using suitable technical standards (such as the Open Archive definitions) that is supported and maintained by an academic institution, scholarly society, government agency, or other well-established organisation that seeks to enable open access, unrestricted distribution, interoperability, and long-term archiving.

Not many institutional repositories fulfil the two conditions of the Berlin Declaration. Most repositories comply with the Open Archives Initiative (OAI) protocol for metadata harvesting, which makes them interoperable. Questions can be raised regarding whether institutional repositories also act in accordance with the conditions regarding long-term archiving or the requested permission statement to use the work freely and widely. Missing this permission statement could have the unwanted consequence that although users can find a work in a repository, they don't know whether they can use the work freely or that they are still bound by the restrictions of copyright.

The open access movement employs two main currents for delivering open access: self-archiving (open access archives or repositories) and publishing in an open access journal. Self-archiving can be described as making e-prints freely available in digital form on the internet. It is referred to as the 'green road', that is, publishing an article in a toll access journal and concurrently archiving it in an institutional open access repository.[15] The most common strategies for self-archiving are depositing an article on the author's personal website, in disciplinary archives, in institutional unit archives or in institutional repositories. This strategy must meet the conditions of open access: the material must be searchable and must fulfil the necessary requirements of a publication, such as the determination of ownership, passing peer review, accessibility and preservation. Open access publishing is the so-called 'golden road'. When an author takes the golden road he publishes his article in an open access journal that makes the articles freely accessible online immediately upon publication. Open access refers to free and unrestricted availability without any further implications. In scientific publishing it is usual to keep an article's content static and to associate it with a fixed author. This is different from the idea of open content which usually is defined to include the general permission to modify a given work. However it is food for lawyers to decide whether the permission statement of the Berlin Declaration refers to open content rather than to open access. Open access is not synonymous with universal access. Even after open access has been achieved, access barriers of censorship, language, handicap and connectivity could hinder accessibility. Suber states nonetheless that there is no reason to hold off using the term open access; 'Removing price and permission barriers is a significant plateau worth recognising with a special name'.[16]

The contents of an institutional repository can be diverse. It may include journal articles, conference papers, e-theses and dissertations, data files and book chapters, whether or not born digitally. More and more authors

also deposit their underlying data and other material regarding their research articles in the repository. Although open access should be immediate, rather than delayed and should apply to full text, not just to abstracts or summaries probably not all the content of repositories is available in full text.[17] Due to restrictions in earlier agreements with publishers it is possible that the full text is only shown campus wide or that only metadata are visible. It could also be possible that provisions of data protection regulations might apply.

The version of a deposited journal article can differ as well. Some of them are preprints, some post-prints or author's versions. DAREnet in the Netherlands tries to acquire the definitive published version. There are several definitions of preprint and post-print in circulation.[18] The copyright toolbox developed by SURF and the JISC employs the definition the Association of Learned & Professional Society Publishers used in its response to the Science and Technological Committee Inquiry to Scientific Publications.[19] The copyright toolbox defines a preprint as a work before it has been peer reviewed, edited or prepared for publication by a publisher.[20] A post-print is a work in the form accepted for publication in which the author has incorporated into the text the outcome of the peer review. The definitive version is the publisher's version that includes further editorial refinement and preparations made by the publisher for producing a version for publication.

The Bethesda Statement states that open access is a property of individual works.[21] This is not necessarily the journal or the publisher involved. It is therefore the copyright owner of an individual work who decides to make his article freely available through one of the two ways of open access. That this is not that obvious showed a recent study in which 32% of the respondents said that the decision on the preferred medium is beyond their control.[22]

Whether or not an author can make his journal article available via the repository depends on the terms and conditions of the agreement that he signed for publishing his article. If they are too strict or when the author has transferred his copyright, distributing the article is liable to the permission of the publisher. For that reason an author must consider carefully whether to sign the publishing agreement unaltered, to amend the publishing agreement to the extend his retains his right to deposit his article in the repository and distribute it through this medium, or to use for instance the Licence to Publish.[23] Depositing a preprint is the author's decision for which he doesn't need to consult his publisher. Depending on the copyright policy of his institution an author can decide himself to archive his preprint. Whether it is possible archive a post print or a definitive version of the journal article the SHERPA database is a reference point.[24] That database gives an overview of publishers' policies about copyright and archiving. Except for using the permission statement as indicated in the Berlin Declaration there are several other means to express the consent to open access. The use of a Creative Commons licence is an easy, effective and

increasingly common way to do so. The part of this study about the author-society relationship will go into this.

Many institutions signed the Berlin Declaration. By doing so an academic institution declares it is in favour of the abstract principle of open access but committing the institution to actually providing open access needs more concrete steps. These steps were formulated at the subsequent Berlin meetings. The meetings provided institutions with a practical open access provision they could commit themselves to after signing the Berlin Declaration. In the beginning of 2005 at the third Berlin meeting the recommendation emerged that in order to implement the Berlin Declaration institutions should: 'Implement a policy to a) require their researchers to deposit a copy of all their published articles in an open access repository and b) encourage their researchers to publish their research articles in open access journals where a suitable journal exists and c) provide the support to enable that to happen'.[25]

It is noticeable that individual universities begin to adapt policies requiring that their researcher employees provide open access. The list of policies of institutional archives in the Registry of OA Repository Material Archiving Policies (ROARMAP) is growing.[26] Also other initiatives concerning the implementation of the Berlin Declaration are taking shape. They are listed at the DRIVER webstie www.driver-community.eu

4.3.2 Role of funding organisations

Many funders have now either made commitments to open access, or are in the process of reviewing their policies and procedures, with a view to opening up access to results of the research they are responsible for. More and more, statements of funding organisations include comments about enhancing access to research publications, especially when such works are financed by public resources and it is often their priority ensuring that the availability and accessibility of the output of research funded by them is not adversely affected by copyright strategies of publishers.

European Research Advisory Board

The European Research Advisory Board (EURAB) recently launched its recommendations to the European Commission on scientific publications and open access.[27] EURAB was invited by the Commission to examine the issue of scientific publications with particular reference to policy recommendations regarding open access for Framework Programme 7. One of the recommendations to the Commission was that as a funding body the European Commission should consider mandating all researchers funded under FP7 to lodge their publications from EC-funded work in an open access repository as soon as possible after publication, to be made openly accessible within six months at the latest.

This recommendation was not fully adopted by the European Commission. In its *Communication to the European Parliament on scientific informa-*

tion in the digital age: access, dissemination and preservation, released the day before the large conference Scientific Publishing in the European Research Area: Access, Dissemination and Preservation in the Digital Age took place, the Commission stated that publicly funded research data should in principle be accessible to all, in line with the 2004 OECD Ministerial Declaration on Access to Research Data from Public Funding, but did not mention a time frame for open accessibility.[28] In this communication the Commission declared that initiatives leading to wider access to and dissemination of scientific information are necessary, especially with regard to journal articles and research data produced on the basis of public funding. In this context, the publications resulting from the research the Commission project costs related to publishing, including Open access publishing, will be eligible for a Community financial contribution. The Commission will encourage the research community to make use of this possibility. The Commission also envisages, within specific programmes (e.g. the programmes managed by the European Research Council), to issue specific guidelines on the publication of articles in open repositories after an embargo period. This would be done on a sectorial basis, taking into account the specificity of the different scholarly and scientific disciplines.

Furthermore the Commission will support research on the scientific publication system within the ERA and globally, for example on publication business models, dissemination strategies, and the connections between research excellence, scientific integrity and the scientific publication system.

Funding organisations as the Wellcome Trust in the United Kingdom or the National Health Institutes (NIH) in the United States insist on specific addenda to insert in publishing agreements that make it possible that the research output is widely distributed. These organisations have extensive policies in which they explain their ideas of open and unrestricted access to published research.[29]

Wellcome Trust

Since 1 October 2006 the Wellcome Trust requires from its grantees that they submit an electronic copy of the final manuscript of their research papers into PubMed central. An important requirement is that the work is not made available to the public later than six months after the official date of the final publication. Additionally the Wellcome Trust expects authors of research papers to maximise the opportunities to make their results available for free and, where possible, to retain their copyright. The organisation has made agreements with several big publishers that will allow authors to comply with the requirements of the Trust.

Research Councils UK

Research Councils UK (RCUK) is a strategic partnership between the eight UK Research Councils. RCUK was established in 2002 to enable the councils to work together more effectively to enhance the overall impact and effectiveness of their research, training and innovation activities, contribut-

ing to the delivery of the UK government's objectives for science and innovation.[30]

In June 2005 the RCUK published a draft position paper on 'access to research output'. The many comments the RCUK received induced the updated position statement in June 2006. The position paper of the RCUK covers all disciplines and covers tow aspects of the changing publication landscape: the author pays publishing and self-archiving.

Concerning the first aspect the RCUK reaffirms its long-standing position that it is the author's choice where to place his research for publication. It is for the author's institution to decide whether it is prepared to use funds for any pay charges or publishing fees. Regarding self-archiving RCUK agrees that their funded researchers should, where required to do so deposit the outputs from research councils funded research in an acceptable repository as designated by the individual research council.

Each individual council gives guidance on the requirement of self-archiving. It will be effective from the time indicated in this guidance. In addition the researcher should wherever possible personally deposit or otherwise ensure the deposit of the bibliographical metadata relating to such articles including a link to the publisher's website at or around the time of the publication. The position statement further emphasises that full implementation of these requirements must be undertaken such that current copyright and licensing policies for example embargo periods or provisions limiting the use of deposited content to non commercial purposes are respected by authors. The research council's position is based on the assumption that publishers will maintain the spirit of their current policies. The individual UK Research Councils have recently released their guidance on open access.

The Arts and Humanities Research Council (ARHC) is committed to the principles articulated in the Research Councils' UK position statement. It planned to finalise amendments by the end of 2006, with a view to ensuring an appropriate deposit of any relevant outputs arising from AHRC applications after then.

The Biotechnology and Biological Sciences Research Council (BBSRC) requires from 1 October 2006 a copy of any resulting published journal article or conference proceedings to be deposited at the earliest opportunity, in an appropriate e-print repository, wherever such a repository is available. Current copyright and licensing policies, such as embargo periods are maintained by publishers and respected by authors.

As AHRC the Engineering and Physical Sciences Research Council (EPSRC) remains strongly committed to the principles outlined in the Research Councils' UK position statement.

The guidance of the Economic and Social Research Council (ESRC) is far more elaborated and states that its funded researchers should deposit the outputs from any research in the ESRC awards and outputs repository where this is permitted by publishers' licensing or copyright arrangements. From 1 October 2006, it is mandatory, at the earliest opportunity, to personally deposit, or otherwise ensure the deposit of, a copy of any resultant arti-

cles published in journals or conference proceedings, in the ESRC awards and outputs repository and wherever possible, personally deposit, or otherwise ensure the deposit of, the bibliographical metadata relating to such articles, including a link to the publisher's website, at or around the time of publication, in the ESRC awards and outputs repository.

Which version of the article should be deposited depends upon publishers' agreements with their authors.

Full implementation of the requirements of ESRC requires that current copyright and licensing policies, such as embargo periods or provisions limiting the use of deposited content to non-commercial purposes, are respected by authors. The ESRC's guidance is based on the assumption that publishers will maintain the spirit of their current policies. Under this policy, at no time will individual authors be required to negotiate copyright and licensing arrangements with their publishers.

The Medical Research Council (MRC) requires for applications submitted from 1 October 2006 that electronic copies of any research papers accepted for publication in a peer-reviewed journal, which are supported in whole or in part by MRC funding, are deposited at the earliest opportunity – and certainly within six months – in UK PubMed Central (UKPMC). The availability of a research paper from PMC (and other PMCI repositories) does not prevent authors from also depositing a copy in their own institutional or another subject-based repository should they choose to do so or be required to do so by their employing institution.

MRC guidance also strongly encourages authors to publish in journals that allow them (or their institutions) to retain ownership of the copyright. The MRC will pay 'author pays' (article processing charges) where these have been included in applications for MRC grant funding.

If the publisher does not permit author/institution-ownership of copyright, authors should publish in journals that permit the paper to be made available in the PMC and PMCI repositories (such as UKPMC) within six months of publication. If a researcher wishes to publish a paper in a journal that is unwilling to agree either to author/institution ownership of copyright, or to allow the article to be made freely available from the PMC and PMCI repositories within six months, the MRC may, in very exceptional cases, grant permission for authors to submit the paper for publication in such a journal. This position will be reviewed in 2008.

Finally from 1 January 2006, all applicants submitting funding proposals to the MRC are expected to include a statement explaining their strategy for data preservation and sharing. MRC data sharing policy indicates that, where possible, published results should provide links to the associated data.

The Natural Environment Research Council (NERC) will establish an e-print repository to improve access to the outputs of its research centres. NERC staff will be expected to deposit copies of any published peer-reviewed papers, supported in whole or in part by NERC-funding, in the NERC repository. NERC award holders who do not have access to an appropriate repository through their host institution will be able to deposit in the

NERC repository. From 1 October 2006 NERC requires that an electronic copy of any published peer-reviewed paper, supported in whole or in part by NERC-funding, is deposited at the earliest opportunity in an e-print repository. To support access to environmental data NERC already requires that award holders offer a copy of any dataset resulting from NERC-funded activities to its data centres. The version of the paper deposited will depend upon publishers' policies on deposit in repositories.

STFC supports the sentiments of the RCUK Councils position statement. For all STFC grants arising from proposals submitted after 1 December 2006, the full text of any articles resulting from the grant that are published in journals or conference proceedings, whether during or after the period of the grant, must be deposited, at the earliest opportunity, in an appropriate e-print repository, wherever such a repository is available, subject to compliance with publisher's copyright and licensing policies. Wherever possible, the article deposited should be the published version. In addition, the bibliographical metadata (including a link to the publisher's web site) must wherever possible be deposited, at or around the time of publication, in the relevant e-print repository.

Nearly all the councils also encourage, but do not formally oblige, all award-holders and staff to ensure deposit of articles arising from grants awarded as a result of applications submitted before 1 October 2006 and most of them will work with publishers to put in place mechanisms for publishers to deposit publications directly, on behalf of authors, where this is possible.

Deutsche Forschungs Gemeinschaft

The Deutsche Forschungs Gemeinschaft (DFG) has tied open access into its funding policy: 'When entering into publishing contracts scientists participating in DFG-funded projects should, as far as possible, permanently reserve a non-exclusive right of exploitation for electronic publication of their research results for the purpose of open access. Here, discipline-specific delay periods of generally six to twelve months can be agreed upon, before which publication of previously published research results in discipline-specific or institutional electronic archives may be prohibited'.

4.3.3 Position of author and his relationships

Author-institution relationship

The copyright owner of a work is the maker of the work. However for scholarly works it is not always unambiguous who the copyright owner of a work is. Although article 7 of the Dutch Copyright Act 1912 states that 'where labour is carried out by an employee consists of the making of certain literary, scientific or artistic works, the employer shall be deemed the author of the work', this didn't stop legal scholars to dispute the appropriateness of this article to scholarly works from the moment of the implementation of the Act in 1912.[31] Nearly a century later it is more or less gen-

erally accepted that the copyright of scholarly works is vested in the author. Also regarding the question in whom the moral rights of a scholarly work are vested there is a difference of opinion; some scholars state that the moral rights are vested in the employer, other scholars adhere to the viewpoint that the maker of the work owns the moral rights.

§43 of the German Copyright Act concerns ownership of works created in the course of employment.[32] In Germany the natural person who created the work has the copyright. A legal person can own a work but an author can give him a licence to exploit the rights of the work in the context of his employment contract. An employee who creates a work as part of the obligations of his employment contract is obliged to grant an implicit licence to his employer as long as the exploitation rights are required for the performance of the contract, even if there is no explicit provision in the contract.

According to the French Intellectual Property Code law (revised by the DADVSI Law, 1 August 2006), the author is the copyright owner. If the author is a civil servant, as scholars are, it is more complicated.[33] Usually, when he acts within the framework of his mission, the administration owns the copyright. However if the publications are not submitted to the control of its hierarchy, the author is considered to be the owner, which is generally the case with a scholarly publication.

In the United Kingdom Section, 11(2) of the Copyright, Designs and Patents Act 1988 (CDPA) states that the author of a work is the first owner of any copyright in it.[34] However when an employee in the course of his employment makes a literary, dramatic, musical or artistic work, his employer is the first owner of any copyright in the work subject to any agreement to the contrary.

Considering the difference of opinion regarding ownership of scholarly works, an author has to find the answer to the question whether he will own the copyright of the resulting scholarly publication before writing a journal article. Could the institution where the research is done and which is employing the author own the copyright or is perhaps the organisation which provides for the funding of the research the copyright holder? The answer to the question of ownership might be found in the copyright policy of the university or research institute that employs the author.

Because of the many debates in relation to ownership of scholarly works, several efforts were made to regulate this effectively. In the late 1990s SURF developed a policy for the Dutch universities which stated that the copyright in academic publications remain vested in the author.[35] In concordance with this policy the author grants the university a licence to use the publication for educational or research purposes without claiming any royalties accruing to him. This policy never landed in the Dutch universities. At that time the institutions were not interested in copyright; patents were thought to be more lucrative.

In 2006 a short study commissioned by SURF and the JISC investigated how universities in the Netherlands and the United Kingdom deal with copyright in terms of their policies and practices especially with respect to the ownership of scholarly works.[36] The study showed that determining

what copyright policy an institution uses was not always easy to find, because often it was not put on paper because it was hard to find. In some cases the copyright policy is part of a wider, sometimes more detailed guidance on IPR in general and patents in particular. Information is sometimes found in more than one place, and not always under the heading of intellectual property. Authors might have to look under 'governance' or 'commercial exploitation and research' to find relevant provisions. Copyright policies can be stated formally or informally in the form of FAQs. In some cases the institution has a staff handbook where information can be found. Libraries are often good sources of information, but in most cases librarians are more focused on third-party material and clearing rights.

The aforementioned study resulted in a list of recommendations to university management on how to adapt their copyright policies and make them more accessible. The study proved once more that the extant legislation, both in the United Kingdom and the Netherlands, seems to be explicit but that the custom and practice found indicate otherwise. Furthermore it showed that not many universities formally deal with the issue of copyright ownership in scholarly works produced by their staff. There often is a pattern of fragmented responsibility for copyright issues within institutions with no clear internal co-ordination. This creates an increasingly complex framework for the establishment of digital repositories.

The study made the following recommendations to universities:

- Copyright needs to be approached as seriously as any other form of IPR. Existing customs and practices should be reviewed with regard to copyright and in particular for scholarly works;
- Clear, official policy on copyright needs to be developed, which also ensures that all employees are aware of this. This policy and supporting information should be disseminated proactively;
- A clear strategy on the ownership and management of copyright has to be used, which takes into account developments in electronic publishing, institutional/digital repositories and the requirements of funding bodies;
- A clear line has to be taken on the assignment or licensing of copyright to publishers by authors of scholarly works in their employ and the implications for re-use and future use of them by the author, his colleagues and the institution as well as the academic community at large. At the same time the academics' freedom to publish has to be upheld;
- The rights stemming from copyright law of the authors need to be supported and upheld, including moral rights, as far as possible;
- Appropriate support, guidance and information on copyright need to be provided to all staff, written in lay terms rather than legal language;
- A person of sufficient seniority needs to be appointed who has to implement the policy on copyright and co-ordinate action on copyright issues;
- A copyright policy should not be developed in isolation but fit into a general approach to copyright in teaching and administrative materials,

software and databases, as part of the whole IPR portfolio of an institution.

Author-publisher relationship

The author-publisher relationship merely determines which rights an author can exercise himself or which he can exercise towards his university. When publishing an article the author agrees with his publisher on the terms and conditions under which his article is going to be published. Therefore an author should identify the rights he may wish to retain. The basic aim here is to create a balance of rights for the stakeholders involved. A publishing agreement can be an important step in achieving this balance of rights and responsibilities in the process of scholarly communication.

Copyright toolbox

The need for balance between the rights of authors, publishers and institutions was the motivation behind for the so-called Zwolle conferences and the Zwolle Group.[37] Zwolle is a small town in the Netherlands where three conferences about the management of rights of scholarly works were held. The Zwolle Group formulated the Zwolle Principles, a set of principles designed for assisting stakeholders to achieve maximum access to scholarship without compromising quality or academic freedom and without denying aspects of costs and rewards involved.[38] The key principles are that the primary focus should be on the allocation of specific rights to various stakeholders (management of copyright) and that optimal management may be achieved through thoughtful development and implementation of policies, contracts and other tools, as well as through processes and educational programs that articulate the allocation of rights and responsibilities with respect to scholarly works.

Building upon the work of the Zwolle Group, SURF initiated the development of a copyright toolbox. This toolbox was compiled in order to enable author and publisher to identify the issues that should be addressed when a scholarly work is being submitted to a journal. The toolbox also provides a publishing agreement that author and publisher can use to set their terms and conditions. In addition the copyright toolbox offers the author sample wording for various options in case he or his publisher would like to amend their agreement. Several pages take the author through a series of provisions designed to assist him in determining which exploitation rights are important to his needs. Linked to the description of each of these rights are portions of text that some publishers and authors have found useful in codifying the involved rights.

Licence to publish

A vital component of the copyright toolbox is the 'Licence to publish'.[39] In this licence SURF identifies the issues to be addressed when submitting an article to a journal whilst at the same time it will be deposited in an institutional repository.

Both the 'Licence to publish' and the sample wording are based on a checklist of key needs[40] which sums up the key needs that are important to author and publisher thereby helping them to determine which rights can be best exercised by which party, thus creating a balance of rights. The interests of authors and publishers often converge, but sometimes they do not. Consideration of both the key needs of authors and publishers helps each of these to understand the other's position when entering into a publishing agreement.

If an author wants to make sure that he retains all the rights needed for optimal access, he can use the 'Licence to publish'. By signing the 'Licence to publish' and sending it to his publisher the author grants the publisher a sole licence, allowing for certain copyright related acts which have an economic or commercial objective with respect to the article. Thus the author retains certain rights for various scholarly purposes, such as depositing the article in a repository. The 'Licence to publish' makes no distinction between preprints, post-prints or author's version but stipulates that the published version of the author's article can be disseminated via an institutional or centralised repository immediately after publication in a journal or after an embargo period of a maximum of six months. The 'Licence to publish' can also be used for multiple authors or joint ownership. One of its clauses deals with this. The 'Licence to publish' is different from other publishing contracts in the respect that the author initiates this contract. The publisher won't have to sign; by accepting the article he subscribes to the conditions of the contract. The 'Licence to publish' is accompanied by 'Principles'.

Digital Peer Publishing Licence

The 'Licence to publish' is a licence concerning publishing in a traditional journal. A contractual basis for publishing e-documents in an e-journal is provided by the Digital Peer Publishing Licence (DDPL).[41] In commission of the Ministry of Science and Research of State of North-Rhine Westphalia Germany, the Institute for Legal Issues on Free and Open Source Software developed and created this licence.[42] DPPL is designed for scholarly content; it covers all aspects of authenticity, citation, bibliographic data and metadata, permanent access and open formats. The licence can be used either by publishers of e-journals or by the authors themselves. The DPPL is modular built, which makes it possible for the licensor to adapt it to his own liking. The licence is customised for national law, it is internationally applicable, and it covers three modules: reading, distributing or accessing verbatim copies, sharing and re-using the work and properly citing if changes are made.

The basic module subjects all documents covered to being read, accessed for downloading and distributed unchanged. No distinction is made between scientific or commercial use. Because this licence only concerns delivery of the document in electronic format, the rights concerning a printed version or a version on storage media are not covered. Thus electronic distribution is promoted. Meanwhile the bearer of the rights retains the option

to close an agreement with a publisher on other versions of his work for commercial distribution. The extended modules of the DPPL address share and reuse of published material. The modular DPPL and the free DPPL allow users to change published material and contain arrangements for proper citing in case changes were made. In the modular DPPL, only those changes that are specifically earmarked may be performed by recipients. This makes it possible, for example, that texts become fixed while images still can be changed. In the free DPPL, everything in the publication remains open to change within the terms of the licence.

Author addenda

Transfer or assignment of all rights to a publisher could lead to loss of control by the author over his scholarly output. Not only will the author no longer be able to re-use his work now and in the future; he also always has to ask his publisher permission for publishing his article in a repository. Institutions and organisations therefore encourage authors to retain their rights. Two ways to achieve this have already been discussed. A third way is to add an author's addendum to a publishing contract. An author's addendum is a standardised legal instrument that modifies the publishing agreement and allows the author to keep his rights.[43] An addendum specifies what rights an author does or does not have in several key areas. An addendum can be attached to a publisher's agreement and is likely to be legally binding. In order to make sure that it is legally binding, it has to be signed by both parties.

Author's Addendum from SPARC

SPARC is an alliance of academic and research libraries and organisations, working to correct market dysfunctions in the scholarly publishing system.[44] The Author's Addendum made by SPARC is a form which an author can use to amend the publishing agreement supplied by a publisher.[45] The Addendum regulates that the author in addition to any right under the publishing agreement retains the right to reproduce, distribute, publicly perform, and publicly display the article in any medium for non-commercial purposes, as well as the right to prepare derivative works and the right to authorise others to make any non-commercial use of the article. The Author's Addendum must be attached to the agreement the publisher has sent to the author and requires of the publisher to demonstrate consent by signing the copy and send it to the author.

Author-society relationship

In the author-society relation the author establishes his relationship with the users of his works. This happens by attaching a permission statement to the work. Copyright holders can either compose their own licence or permission statement or use one of the many open content licences. The use of one of the Creative Commons licences is an easy, effective, and increasingly common way to make clear to society how a work can be used.

It is important to state how a work can be used. If a work does not carry a licence then the normal rules of copyright apply. This means that the exceptions and limitations of the law determine the use an end user can make of the work. This use is far more limited than the free, irrevocable worldwide right of access granted under the permission statement of the Berlin Declaration.

The author-society relation also covers the relation between author and repository. The author and his institution need to make specific arrangements about the works an author is going to deposit in the institutional repository. These arrangements should be considered an essential part of a digital repository operation. They establish obligations and rights of both parties in a formal way. For each deposit into the repository the author has to give permission to the repository, firstly to store and preserve the work, and secondly to make it available under set conditions, the latter depending on the existence of a publishing agreement and its terms and conditions. This means that there should be a deposit licence between the author and the institution. Just as a publishing agreement sets the rights and obligations for publishing an article, so does a deposit licence set the conditions for preserving and making available scholarly works. The SHERPA report on deposit licences for e-prints[46]indicates that the majority of deposit licences cover four topics: the ability of the depositor to legally deposit the e-print, the rights the depositor maintains over the deposited work, the permissions the repository gains to maintain the deposited work and the conditions under which the repository can remove the e-print.

Deposit licence

A deposit licence must address several rights and obligations. The author has to give permission to the repository to store, reproduce and migrate the work in order to keep it available and accessible, irrespective of form or medium. A digital repository has an important role in safeguarding permanent access to the deposited material. Therefore, the archival function of the repository needs to get sufficient attention. In addition, the author has to give permission to distribute the work and make it available by transmission on line or in any other form. Finally the repository manager or the institution must have permission to make the work available to users under a non-exclusive irrevocable licence that allows the user to reproduce and distribute the work in any medium en in every format under the condition of attribution.

The obligations of the repository need to be written out clearly. It should attribute the work to the author and state to the users that they are obliged to give proper attribution when using the work. The repository must archive the work permanently, and keep it readable and accessible. If there is an embargo period before the work can be made available full text, the repository manager must provide adequate technical protection measures to prevent that unqualified users gain access to the material.

It can be argued that a deposit licence should not contain a provision that makes it possible to exploit the work commercially. For reasons for sustain-

ability and when creating a layer of services on a digital repository, an institution can be interested in having the possibilities for commercial exploitation. Because intellectual property is a property, an institution cannot mandate in its copyright policy that an author has to give up his exploitation rights via a deposit licence. Therefore a carefully drafted mandatory deposit licence cannot contain provisions that deprive an author of his right to decide himself how to exploit his works.

Legal toolkit Leiden University

Many electronic deposit licences have been developed. An overview is given in the 'Legal Toolkit', a project funded by SURF and conducted by Leiden University in the Netherlands.[47] This toolkit offers guidance concerning the various e-deposit licences used around the world, and gives recommendations for use.

Leiden University made a list of recommendations also for setting up a legal framework for deposit licences for theses and dissertations. They recommended that an embargo on making the text fully available should only be established if there are ponderous reasons. These could involve publishing rights, confidential information or requests for patents. Embargoes should be established in close cooperation with the author of the work. Another recommendation made by Leiden was that the duration of the deposit licence should be unlimited and irrevocable. In addition commercial use should only be allowed if the author would get a reasonable remuneration. They also recommended creating an archive for deposit licences. Because of the nature of rights that are ascertained in the deposit licence it is recommendable to give out these rights via the OAI-PMH interface for end users and harvesters.

Licence to deposit

In line with the Licence to publish SURF has drafted a 'Licence to deposit'. SURF also accompanied this Licence to deposit with a set of 'Principles to deposit'. An author might use these principles when depositing his published work, including accompanying data, models or visualisations, in the digital repository. As in the Licence to publish the author retains his copyright. Via the Licence to deposit he gives the owner of the digital repository permission to store his work and keep it permanently accessible. The repository can make the work available under the conditions of the Berlin Declaration. The Licence to deposit comes into effect after transfer of the work to the repository. The 'Licence to deposit' is irrevocable, and the stored publications stay in the repository. Access to the work can be denied only for ponderous reasons. An embargo period of six months is optional.

Creative Commons

The use of Creative Commons licences in higher education is disputed. Some authors consider certain Creative Commons licences as very suitable for distributing a scholarly article,[48] others give a number of reasons why institutions of higher education should think twice about using these li-

cences.[49] However, if there is a licence attached to material in a repository of scholarly works, more often than not it is a Creative Commons licence. The Creative Commons licence was developed at Stanford University in 2001. In the core licensing suite is a total of six licences to choose from, each of which each permits different uses of the work. They are expressed in three different ways: a plain explanation of the licence together with the relevant icons that indicate the scope of the permitted use, the legal document and the machine readable code.

The most liberal of the core suite Creative Commons licences is the attribution (by) licence. Under this licence users are permitted to copy, distribute, display, and build upon the author's work as long as they name the original maker of the work. Under this licence a user can make use of the work commercially.

A work with the Attribution Share Alike licence (by-sa) attached to it can be copied, distributed, displayed, and performed as long as the newly created work is licensed under identical terms. A user must attribute the original author and can use the work commercially if he wants to.

The Attribution No Derivatives (by-nd) allows a user to redistribute a verbatim copy of the work commercially and non-commercially under acknowledgment of the creator.

A user can copy, distribute, display and perform a work non-commercially under an Attribution Non-Commercial (by-nc) licence. The Attribution Non-Commercial licence authorises others to copy, distribute, display and perform the work, and derivative works based upon it- but for non-commercial purposes only. The new work must bear the name of the author but it does not have to be distributed under the same terms and conditions.

The Attribution Non-Commercial Share Alike (by-nc-sa) licence allows users to remix, tweak, and build upon a work non-commercially with acknowledgement and further licensing under the same terms. Other users can download and distribute the work. Furthermore they can translate, remix and tweak. All new works will carry the same licence and derivatives will be non-commercial.

The strictest licence is the Attribution Non-commercial Non-Derivatives (by-nc-nd) licence. Under this licence users can redistribute the work under attribution of the original author. This licence does not allow for the work to be changed in any way.

An institution needs to be aware that it cannot attach a Creative Commons licence to a work in its repository without the consent of the copyright owner. It is always the copyright owner who decides under which conditions his work can be re-used. In case of a digital repository the institution is acting as an intermediary. It can be argued that in this case the repository is also a user and therefore is authorised to exercise the permitted rights. This would mean that no e-deposit licence is needed for the distribution of works, because that is clearly written out in a Creative Commons licence. However, from a managerial and risk-avoiding point of view, an academic institution should perhaps like to specify some rights not cov-

ered in the Creative Commons licence. For instance, the Creative Commons licence does not oblige that the repository guarantees the availability of the work in the future or preserves the work digitally.

5. Data curation

René van Horik

5.1 Introduction

Scientific research increasingly creates and uses digital data in many different ways and in a wide range of formats. Data curation activities are required to maintain and preserve the digital research data as well as to facilitate its future reuse. Increasingly science will be carried out through distributed global collaborations enabled by the internet. The Grid infrastructure will provide 'e-Science' with powerful large-scale computing resources and dedicated repository management software.

This chapter on data curation is closely related to chapter 6 by Barbara Sierman on digital preservation, as both studies are concerned with the longevity and long-term storage of digital objects. The main difference between the two chapters is that her chapter on digital preservation takes organisation and management aspects into consideration whereas this chapter on data curation takes the digital object as its starting point. Despite the fact that the two reports emphasise different aspects, some overlap is unavoidable.

This chapter consists of five parts. The first part contains an elaboration on the concept of data curation. Next, features of scientific digital objects are described for which data curation is relevant. The third part covers data quality issues. Next, some remarks are made on data curation tools and procedures; the tools, procedures and concepts that are described in this section are examples of practical implementations of data curation. The last part of this chapter contains concluding remarks. As data curation is a relatively new term and used within the context of a wide number of projects, initiatives and organisations in different ways, it is impossible to cover all aspects and details in an objective way. This chapter is based on a number of published information sources and empirical observations.

5.2 What is data curation?

According to Wikipedia, a curator is a person in charge of a cultural heritage institute (e.g. an archive, gallery, library, museum) who cares for the institution's collections. The object of a curator's concern necessarily involves tangible objects of some sort, whether it is artwork, collectibles, historic items or scientific collections. The role of the curator encompasses collecting objects, making provision for the effective preservation, conser-

vation, interpretation, documentation, research and display of the collection, and to make the collection accessible to the public.[1]

Increasingly curators are active in the digital domain. Digital curation or data curation is needed to maintain digital materials, such as research data, over their entire life cycle and over time for current and future generations of users. Data curation is closely related to digital preservation as both activities are aimed at long-term storage, access, and usage of digital objects.[2] Often the terms are used interchangeably, but there are some subtle differences between the two. Curation not only implies the preservation and maintenance of a collection or database, but also relates to the creation of added value and knowledge. The differences between digital preservation and data curation are caused by the fact that both concepts have their roots in different user communities. The scientific community primarily uses 'data curation', whereas 'digital preservation' has its roots in the digital library community. Data curation starts more or less bottom-up and is increasingly a point of attention for scientists and organisations that support scientific activities and that are using information technology such as scientific data archives.

At the end of the last century archives and libraries found that the objects in their collections were increasingly becoming digital. Based on existing concepts of analogue archival records and analogue publications, solutions were sought to cope with the selection, appraisal, storage and dissemination of their digital counterparts. We will see further on in this study that both the digital library community and the organisations supporting scholarly communication are constructing their own distinctive conceptual model of digital scientific objects. Currently there is an extensive exchange of ideas between the two groups.

Important platforms where conceptual models on data curation are constructed and implemented, and where libraries, archives, museums, scientific data centres, IT research centre and other producers and publishers of scientific data exchange ideas, are the EU-funded project Planets, the project Caspar, the digital preservation cluster of the Delos network and the DPE project. Also important are European research infrastructures. Examples of these are the digital research infrastructure for the arts and humanities DARIAH, the CLARIN initiative for a research infrastructure on language resources, the European research observatory for the humanities and social sciences EROHS, and the CESSDA research infrastructure for the social science data archives.[3]

Data curation is a relatively new term. According to Beagrie, the term was used for the first time with its current understanding at a seminar in London organised by the Digital Preservation Coalition in 2001.[4] A strong indicator that the data curation concept is gaining ground is the fact that in 2006 the peer-reviewed *International Journal of Digital Curation* was founded, of course in line with the spirit of the times as an open access digital journal.[5] In this journal Beagrie explores the emerging field of digital curation as an area of inter-disciplinary research and practice. The selection and maintenance of a body of knowledge and evidence for specific dis-

ciplines or topics are important curation activities. Issues involved here are annotation, linkage, management, validation and editorial input of domain specialists. The archiving and preservation of digital research data is not an end-of-project activity, but has a connection to the creation of these materials and the promotion of its re-use. A life cycle approach to the maintenance of digital research data is important. For this, different (and often differently interested) stakeholders must become involved with data resources at different stages.[6]

In another publication Beagrie elaborates on the difference between digital preservation and digital curation. He states that digital preservation has been used for the series of managed activities necessary to address preservation challenges and to ensure continued access to digital information as long as necessary. Alongside digital preservation the term digital curation is being used for the actions needed to add value and to maintain digital research assets over time for current and future generations of users. The concepts of digital preservation and curation, Beagrie writes, are still relatively new and usage varies between sectors and disciplines but they should be seen as closely integrated and complementary terms.[7]

Increasingly, scientific publications will be available in digital form. The volume of other academic data objects will also increase dramatically. These might include data generated from sensors, satellites, computer simulations, high-throughput devices, scientific images, digital capture devices, and the like. An example is the Large Hadron Collider (LHC) at CERN (Geneva) that will generate roughly 15 petabytes of data annually from 2007, which thousands of scientists from around the world will access and analyse. Digital scientific resources are growing in volume and complexity at a staggering rate. The cost of producing the resources is very high, thereby justifying the attention for data curation.

According to Beagrie the funding for repositories of scientific data is unlikely to match the exponential growth in data and publications currently underway.[8] There will be a need for more automation of processes and metadata generation, software tools for this, and potentially the development of greater collaboration and shared services to lower the entry and operational costs for institutions. Not all of the digital information will have long-term value. So selection for long-term curation will be a significant issue. As a consequence, selection, curation, and long-term preservation of digital resources could be of increasing importance.

The implementation of data curation requires new skills.[9] In some subjects, databases are supplementing or partly replacing journal publications as a medium of scholarly communication. Some have dedicated curators, but some are too small and project based. The web as an electronic publication medium involves increasingly dynamic, on the fly generation rather than static fixed versions of content. As the volume, complexity, and heterogeneity of digital information grows, the requirement for active management becomes more challenging and more critical to a wider range of organisations. This is not only a technical issue. It also involves social factors and organisational risks particularly over extended periods of time.

Digital scientific knowledge, if it is to be useful and useable, must be continuously updated, maintained and accessed. The emerging field of digital curation is central to this process. Persistent information infrastructures for digital materials must be developed. Digital curation skills of researchers and information professionals must be developed.

5.3 Digital scientific objects

This section discusses the features of the digital scientific objects for which data curation is relevant. In the recent past new types of academic digital objects emerged and in this section an attempt is made to identify these objects. It is not easy to assess digital scientific objects in an unambiguous way as a number of theories, conceptual models and perceptions do exist in this field. IT innovations enabled the emergence of these new types of digital objects and new ways of scholarly communication.

The OAIS reference model establishes a framework of terms and concepts relevant for the long-term archiving of any type of digital data, but the features of these digital data are not provided by the OAIS standard.[10] As long as the so-called 'designated community' can understand the information packages that are attached to the digital objects, the requirements of the OAIS standard are met. Currently a number of certification processes are underway and the outcomes of these projects will contribute to the establishment of canonised terms and the understanding of the features of digital objects relevant for the scientific community.[11]

The characteristics of the scientific data objects determine to a large extent the required data curation activities. Libraries have a long tradition regarding the creation of documentation or metadata for information entities such as printed books and journals. In the course of time the archives and libraries communities developed several bibliographic languages. The computer revolution changed the nature of entities to be organised and the means of their organisation and the existing bibliographic languages were adapted to the new situation. One of the problems here relates to the nature of digital objects. A traditional document, like a book, tends to be correspondent to a discrete physical object. On the other hand a digital object can be unstable, dynamic and without boundaries. What is difficult to identify is difficult to describe and therefore difficult to organise and, adapted from Svenonius, difficult to curate.[12]

5.3.1 Typology of digital scientific objects

A wide range of descriptions can be found of scientific data objects that are stored in a digital data repository. Tindemans, for example, uses the term 'record of science' for data collected by scientists and scholars in experiments, observations, surveys, simulations, and databases of historical or sociological events.[13] Heery and Anderson, in their review of digital reposi-

tories, make a distinction between eight kinds of data objects based on their content type.[14] These are raw research data, derived research data, full-text preprint scholarly papers, full-text peer-reviewed final drafts of journal/conference proceedings papers, e-theses, full-text original publications, learning objects, and corporate records. Other typologies applied by Heery and Anderson are coverage, functionality and target user group.

Brogan modified and abbreviated the repository typology of Heery and Anderson. This typology is given in the table below;

Via content type:
- Research data
- Research output
- E-theses
- Learning materials
- Multimedia
- Assessment materials
- Corporate records

Via primary functionality:
- Subject access to resources
- Enhanced access to resources
- Preservation of digital resources
- New modes of dissemination / publication
- Sharing and reuse of resources

Via coverage:
- Personal / informal
- Journal
- Institutional / departmental
- Inter-institutional
- National
- Geospatial

Via target user group:
- Learners
- Teachers
- Researchers

Figure 9 – Repository typology[15]

Despite the fact that a lot of publications, research and projects acknowledge that next to the ubiquitous genre of the digital publication a wide range of new types of digital objects came into existence there is no consensus on the characteristics, names and definitions of these new types of digital objects. Most of the current digital repositories maintained by the archive and library community contain digital publications, such as scientific journals and dissertations. This is confirmed by the results of the survey on repositories in Europe that is part of the DRIVER project.

One could even argue that repositories with multimedia objects, such as digital still images and digital moving images, are basically document-oriented repositories containing 'non-book' publications. The management and long-term archiving policies of these multimedia objects very much resemble the digital publications approach. The main difference between the two is that each group of objects is processed with dedicated metadata schemes, or, in the case that a common metadata schema is used (e.g. the Dublin Core metadata element set), the elements are interpreted according to the features of the scientific digital object.

The current 'publication'-oriented character of trusted digital repositories that provide access to scientific output confirms that in practice data curation is limited to a specific, one could say traditional, type of scientific digi-

tal object. An example of this is the DARE distributed repository that provides access to more than 150,000 academic publications such as digital scientific journal articles and electronic theses.[16]

Besides the construction of a distributed repository of publications, the DARE project also gave the initial impetus to data curation solutions for other types of scientific digital objects. Examples are the DARELUX project aimed at the durable storage and access of hydrological data, the EDNA project aimed at the durable archiving of digital data on archaeological research and the DARC project aimed at providing access to African studies research material and information accessible through a community portal on the internet. The DARE project E-laborate created a proof-of-concept digital repository of scholarly literary sources with additional data curation functions that facilitate the creation of added value and knowledge. E-Laborate is intended as a virtual workplace for researchers in the humanities and social sciences. The four projects mentioned above illustrate that only the first steps have been made towards a general consensus on what kind of scientific digital objects can be distinguished and how the curation of these objects should be implemented.[17]

Currently the publication-like data object has a predominant position, while a generally accepted notion of scientific data objects in general is lacking. A way to improve this situation is to design a conceptual model that contains all essential features of scientific data objects in line with the way scientific communication is organised. In contrast with the 'traditional' scientific publication, Hunter emphasises the importance of workflow and lineage in the model for scientific data objects. Workflow technologies represent an increasingly important component of the scientific process. They capture the chain (or pipeline) of processing steps used to generate scientific data and derived products. They also enable scientists to describe and carry out their experimental processes in a repeatable, verifiable and distributed way and to track the source of errors, anomalies or faulty processing.[18]

A number of conceptual modelling initiatives can be distinguished that are relevant for the establishment of consensus on the features of digital scientific objects that must be curated. As an example two of these conceptual models are described in more detail: the first model – the ABC model – is more or less founded in the digital library community, whereas the second one – the CCLRC Scientific Metadata Model (CSMDM) – bases its universe of discourse on specific scientific disciplines, mainly in the sciences.

5.3.2 The extended ABC model

The ABC model is designed to enable the precise recording of life cycle events for digital objects in library, archives and museum domains.[19] Figure 10 contains a graphical representation of the ABC model.[20]

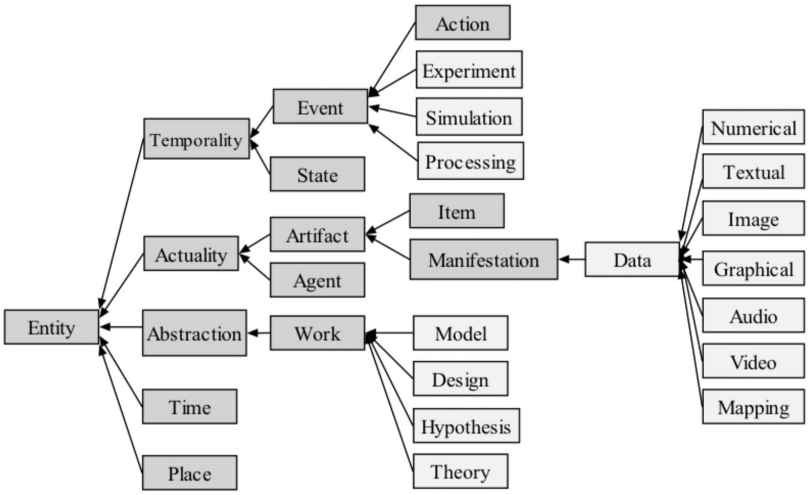

Figure 10 – The extended ABC model class hierarchy[21]

Particularly for data curation, the temporality class is of relevance. The ABC model facilitates the way in which properties of objects are transformed over time. The ABC model makes it possible to unambiguously express situations in which object properties exist, the transitions that demarcate those situations, and the actions and agencies that participate in those transitions.

The model also refers to a very important standard firmly grounded in the traditional library community, which is the IFLA report on functional requirements for bibliographic records, often abbreviated as FRBR.[22] The IFLA FRBR, also called the 'conceptual model for the bibliographic universe', is based around an object of intellectual content called a 'Work'. The concepts 'Item' and 'Expression' also originate from the IFLA FRBR.

Hunter extends the ABC model in order to capture the provenance or lineage of scientific output.[23] New subclasses are associated with the existing IFLA FRBR classes. This extension of the ABC model is relevant for data curation because it makes scientific digital objects more clear. In the extended ABC model the 'Data' class (which is a subclass of the 'Manifestation' class) has the following subclasses: 'Numerical', 'Textual', 'Image', 'Graphical', 'Audio', 'Video' and 'Mapping'. The 'Work' class gets the subclasses 'Model', 'Design', 'Hypothesis' and 'Theory'. The 'Event' class has the subclasses 'Experiment', 'Simulation' and 'Processing'. It is beyond the scope of this study to assess to what extent the extended ABC model by Hunter can handle real-life situations.

Based on an extension of the ABC model, Hunter introduces the 'Scientific Publication Package' (SPP) as a new information format that encapsulates raw data, derived products, algorithms, software, textual publications, and associated contextual and provenance metadata.[24] This new information format is fundamentally different from the traditional file-based for-

mats as they are known by library researchers. Hunter describes a high-level architecture that is currently under development that enables scientists to capture index, store, share, exchange, re-use, compare and integrate scientific results through SPPs. As such this architecture is a very good example of a data curation system. The ABC model and the SPP are based on a number of scientific concept models for publishing scientific data and results and for documenting the lineage of scientific theories and advances.[25]

Hunter stresses the importance of workflow technologies as a component of the scientific process. They capture the chain of processing steps used to generate scientific data and derived products. They also enable scientists to describe and carry out their experimental processes in a repeatable, verifiable and distributed way and to track the source of errors, anomalies or faulty processing. Consequently, a number of international research groups are concentrating on developing workflow specification and enactment systems that allow scientists to easily define, save, edit, share and re-use their workflows.[26]

5.3.3 The CCLRC Scientific Metadata Model (CSMDM)[27]

The CCLRC model attempts to provide a generic metadata model to describe scientific data holdings from the perspective of so-called 'Studies'. The CCLRC Scientific Metadata Model (CSMDM) provides a high-level generic model, which can be customised to specific scientific disciplines.[28] The data model attempts to capture scientific activities at different levels: at the top level there are 'Policies' which are enacted by initiating and maintaining 'Programmes' that consist of one or more generic activities called 'Studies'. Each 'Study' has one or more 'Investigations' that can be of different types (e.g. 'Measurement', 'Simulation', 'Experiment', etc). Figure 11 contains a graphic representation of the model.

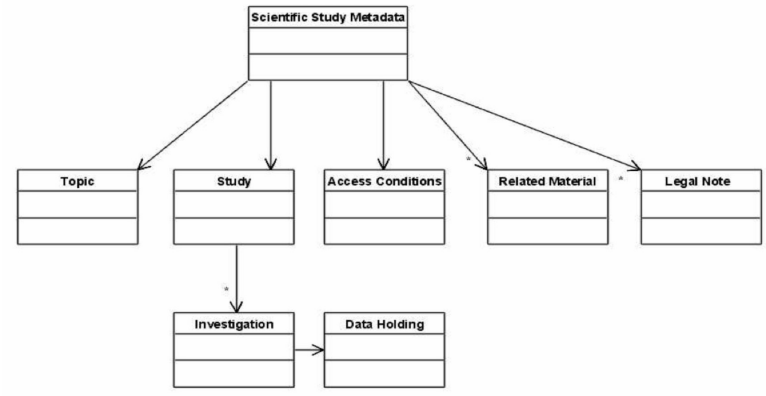

Figure 11 – Scientific study metadata hierarchy[29]

Compared to the extended ABC model, the CSMDM model is less dynamic and not able to express the temporal dimension of digital research data. Another observation that can be made is that the model, which has its foundations in the library community (the IFLA FRBR model), has more advanced features than the science-based CSMDM model. The latter very much resembles a traditional catalogue of more or less fixed objects. The 'Investigation' class probably makes it possible to re-execute experiments.

The most important conclusion that can be drawn from the two conceptual models and definitions discussed above is that currently a number of communities are engaged in fixing and disseminating the concepts, classes and models that are relevant for data curation. It should be stressed that the are only two examples of the work that has been done to elucidate the features of scientific data objects and the kind of infrastructure that is required to safely archive, re-use, refer to, gain credits and authenticate them. The projects and initiatives mentioned in footnote 139 are very important with respect to the development of data curation models and solutions.

Despite the fact that a clear and general accepted definition of scientific data objects does not exist and probably never will, in the next section a number of observations are made concerning the tasks and functions related to data curation.

5.4 Data curation and data quality

Data curation implies the care of research data in a managed environment by dedicated organisations such as scientific data archives. A number of data quality issues can be considered as relevant for data curation activities.[30] The quality of digital research data should meet the following five conditions:

1. Digital research data must be *findable* by means of a catalogue on the internet. This makes appropriate documentation of the research data relevant.
2. Digital research data must be *accessible*, provided that privacy rules and intellectual property conditions are taken into consideration. The ultimate goal is to realise open access to the research data.[31]
3. Digital research data must be available in a *useable* data format, enabling secondary analysis in the future. Therefore the research data must be available in a format that can be processed by common available hardware and software, now and in the future.
4. Digital research data must be *reliable*, that is, the research data is authentic and not changed in the course of time.
5. Digital research data must be *referable* in a durable manner. This implies that the research data is provided with persistent identifiers and stored in a in a so-called trusted digital repository.

Based on the five quality conditions mentioned above practical recommendations can be formulated. These recommendations are relevant in the fol-

lowing three quality areas: 1) the *quality of the data* format and data content, 2) the *quality of the usage* of the digital research data and 3) the *quality of the data storage-facility* or repository. In the next section, the three quality areas are described in more detail.

5.4.1 Quality of the data content and data format

It is evident that the creator of the research data is always responsible and accountable in every circumstance for the scientific content of the material. Fellow researchers and other experts must be able to review and assess the quality of the research data. The common peer-review process that is used for the determination of the scientific value of a publication in a scientific journal can be used as a model for the determination of the scientific value of scientific research data.

A review procedure for research data should enable the answering of a number of questions concerning the quality, value and background of the research data. Examples of these questions are: 1) Are the research data useful for specific research and available for re-use? 2) Are the research data based on original work of the depositor and does the depositor have a good name? and 3) Is the collection of the research data carried out according to common accepted procedures in the given scientific discipline?

The data archive that preserves the scientific data is not able to judge the scientific value and scientific quality of the research data. What the data archive should do, however, is to facilitate the judgement of research data by external scholars and other stakeholders. For this, detailed metadata of the research data are required as well as easy access to the research data that is described with this metadata. The research data review facility should enable a number of judgements on the research data, such as: 1) a judgement on the possibility to re-use the research data in other research fields or by an other research community; 2) a judgement on quality of the researcher or research group related to the research data; and 3) a judgement on the data collection method applied by the researchers.

A large number of data formats do exist that can be used to structure the different types of digital research data.[32] There is a risk that data formats will become obsolete resulting in unusable data objects. For this reason, lasting standardised data formats should be used. As part of the data curation activities it is the task of the data archive to monitor the state of art technologies of data formats and to take active measures to guarantee that research data stays accessible in the long run. Migration of data objects to new formats is an example of these activities.

The quality of the metadata contributes to the easy access to the research data. Three types of metadata can be distinguished: descriptive metadata, structural metadata and administrative metadata. The function of descriptive metadata is to gain insight in the relevance and meaning of the research data. Descriptive metadata is also used to retrieve a given data set from the data archive.[33] The function of structural metadata is to address

the relation between parts of a scientific data set. Structural metadata is required in order to be able to process the research data and to gain insight in its components. The administrative metadata facilitates the access to the research data. Examples of types of administrative metadata are a description of the intellectual property rights, licensing information and the preservation metadata that is required for the durable archiving of the research data.[34]

5.4.2. Quality of the data usage

The quality of the usage of the research data is determined to a large extent by the absence of obstructions to get access to the research data. Scientific research benefits from the absence of access barriers to research data. The principle of open access to scientific data is gaining ground. The open access principle is based on the idea that all scientific data should be easy and freely accessible for everyone.[35] Open access means that users of research data have permission to read, analyse, download, copy, distribute, print, and to link the data without any financial, legal, or technical obstruction. The only, but very important, requirement that must be met is that in all circumstances, the creators of the research data retain control of the integrity of the data. The work has to be cited in the correct way and the creator of the research data must be acknowledged.

At the very least, the metadata of the research data should be freely accessible. Part of the metadata is the conditions that must be met in order to gain access to the research data. These conditions are a component of the user licenses related to the research data. The establishment of the license conditions is part of the transfer process of the research data from the creator to the data archive and are incorporated in the administrative metadata mentioned above.

Next to open access the quality of data usage is influenced by a number of codes of conduct. The fair and lawful use and security of personal data is monitored by a number of organisations and the expertise of these organisations must be used in situations where personal data is part of the deposited scientific data.[36] The legal use of personal data is a responsibility of the depositor of the research data. The legal limitations for the use of personal data should be incorporated in the license conditions that are part of the open access regulation.

Other important codes of conduct relevant for the quality of the usage of research data are those in the field of the scientific practices. In the Netherlands, for example, the Association of Universities in the Netherlands established 'The Netherlands code of conduct for scientific practice: Principles of good scientific teaching and research'.[37] This code is based on five principles. The first principle, *scrupulousness*, states that scientific activities are performed scrupulously and should be unaffected by mounting pressure to achieve. The second principle, *reliability*, is based on the observation that science's reputation of reliability is confirmed and enhanced through

the conduct of every scientific practitioner. A scientific practitioner is reliable in the performance of his/her research and in the reporting, and equally in the transfer of knowledge through teaching and publication. The next principle, *verifiability*, is described as follows. Presented information is verifiable. Whenever research results are publicised, it is made clear what the data and the conclusions are based on, where they were derived from and how they can be verified. The fourth principle of good scientific teaching and research is *impartiality*. This implies that the scientific practitioner heeds no other interest apart from the scientific interest. In this respect, they are always prepared to account for their actions. The fifth and last principle, *independence*, states that scientific practitioners operate in a context of academic liberty and independence. Insofar as restrictions of that liberty are inevitable, these are clearly stated.

The scrupulousness, reliability and verifiability principles are closely related to data curation issues. Scrupulousness is an important principle because accurate source referencing and precise publishing of the research results are examples of best practices in data curation. A best practice of the reliability principle refers to the system of peer review that, as we have seen, is an important component of data curation. To conclude, the verifiability principle contains a number of best practices that are relevant for data curation. The verifiability principle implies that the quality of data collection, data input, data storage and data processing is guarded closely. All steps must be properly reported and their execution must be properly monitored. Raw research data are archived in such a way that they can be consulted at minimum expense of time and effort.

5.4.3 Quality of the data storage-facility

The data storage facility is responsible for the long-term maintenance and preservation of the research data. An important document containing recommendations concerning the long-term preservation of digital research data is the 'Memorandum on the long-term accessibility of digital information', formulated by the German Nestor project.[38] The quality of the data storage facility of the research data is determined by the quality of its organisational framework and quality of the technical infrastructure. Often the concept 'trusted digital repository' is used in relation to the organisational and technical tasks aimed at providing long-term access to digital data sources for a designated community.[39]

Scientific data archives that implement digital archiving tasks and that create trusted digital repositories must have a sound financial, legal and organisational basis, also in the long run. Some organisational characteristics of a scientific data archive are:

1. The data archive has an explicit mission in the field of data curation and data archiving and disseminates this mission.
2. The data archive facilitates the optimal usage of the research data for external users.

3. The data archive implements and complies with all relevant legal regulations and contracts.
4. The data archive carries out an active quality management, based on the principles described in this section of the report.

Concerning the quality of the scientific data objects that are part of the holdings of the scientific data archive the following organisational remarks can be given:

1. The data archive guarantees the integrity of the data objects as well as its metadata during all processing phases.
2. The data archive guarantees the authenticity of the data objects as well as its metadata during all processing phases.
3. The data archive implements a long-term planning of measurements relevant for durable archiving.
4. The data archive adopts research data from data producers.
5. The data archiving activities are executed according to criteria that are settled in advance.
6. The data archive removes all barriers for actual usage of the data objects.

Technical tools, instruments and procedures are required to implement the organisational requirements as described above. Some of these tools are available and others are under development. A number of these technical tools are described in the next section. These tools are often created by international groups of stakeholders active in the field of digital preservation. It is important that data archives and other organisations that are responsible for the curation of scientific data objects monitor and participate in these activities.

To conclude, it can be stated that data curation is the responsibility of a number of organisations, such as funding bodies, research organisations and service organisations. Funding bodies should stress the importance of data curation in order to avoid the loss of investments. The organisations that carry out the research and create the research data should take the future reuse of the data into consideration and e.g. put effort in the creation of high quality metadata and the application of durable data formats. Service organisations that facilitate the optimal storage and usage of digital research data are in first instance responsible for the practical implementation of data curation.

5.5 Data curation tools and procedures

We have seen that the heterogeneous nature of scientific digital data objects as well as its wide range of formats and data types complicates the unambiguous formulation and implementation of data curation tasks and functions. In this section a number of practical tools and services related to data curation are covered. After all, it is relevant to discuss components that are available to implement some of the principles discussed in the earlier sec-

tions of this report. It should be noted that there is currently no general consensus on how to implement and execute data curation. A number of initiatives are taken towards the establishment of procedures, tools and standards relevant for data curation. Some of the most important data curation tools, standards and procedures are described in this section.

An important source of information concerning tools, standards and procedures for digital preservation and data curation are the 'curation manuals' of the Digital Curation Centre (DCC).[40] The DCC is an initiative to provide a range of support services on digital curation and preservation, and to conduct research in this area. The DCC Digital Curation Manual is a community-driven resource – from the selection of instalment topics to authorship and peer review.[41]

Both the reports 'E-infrastructure strategy for research' (Beagrie, 2007) and 'Mind the gap: assessing digital preservation needs in the UK' (Waller and Sharpe, 2006) state that too few tools have been created to help organisations perform digital preservation activities such as performing format migrations, format validation and automated metadata extraction.[42]

Data curation tools are used to implement services that are aimed at creating added value for digital objects deposited at a service organisation. Enrichment of these digital objects, for example by creating metadata, is an example of added value of the digital objects. In the remaining part of this chapter a number of tools, procedures and concepts are discussed that can play a role in the practical implementation of the data curation of scientific digital data objects.

5.5.1 Digital data format registration tools

The way the binary digits are arranged in a digital file depends on the file format specifications. Information on the internal syntax and semantics of the file format is important in order to understand and process the digital file. Format registries that contain representation information about digital formats can help to ensure long-term access to digital files. A format registry can be used to identify, validate, characterise, transform and deliver digital objects, even in the long run.

The Global Digital Format Registry (GDFR) is an example of an initiative that investigates the possibilities to establish a sustainable data format registry.[43] The data model design of the registry system was driven by consideration of the question: 'What information would you want to have today to deal with a digital artefact from 50 years ago?'[44] A proof-of-concept prototype of the GDFR is under development, but it is still far from being an operational production registry.

The National Archives in the UK started a file format registry under the name PRONOM.[45] As stated on its website 'PRONOM is a resource for anyone requiring impartial and definitive information about the file formats, software products and other technical components required to sup-

port long-term access to electronic records and other digital objects of cultural, historical or business value'.

5.5.2 Tools for digital data object identification, validation and characterisation

Besides format registries, tools have been developed to perform format related identification, validation and characterisation of digital objects. Identification is the process of determining the specific format of a digital object. Validation is the process of determining the conformance of a digital object to the specifications for its purported format. Characterisation is the process of extracting preservation information or metadata from an object.

JHOVE (JSTOR/Harvard Object Validation Environment) is an extensible framework for the format-related identification, validation and characterisation of digital objects.[46] The JHOVE programme currently available contains 12 modules, such as audio file formats (AIFF and WAVE), ASCII, digital image file formats (JPEG, GIF, TIFF), PDF and the mark-up languages XML and HTML.

5.5.3 Automatic metadata extraction and metadata registries

It is obvious that, in order to create added value of digital objects, it is important to have detailed information on the features of the digital objects. Metadata helps to assess this value. The creation of metadata can be done manually by means of the entry of subject headings, keywords and other descriptors in a catalogue, but this is a very labour-intensive activity. Automatic procedures can make this process much more efficient.

The National Library of New Zealand has developed a software tool to extract preservation metadata from the headers of a range of file formats. The preservation metadata extract tool automatically extracts preservation-related metadata from digital files, and outputs that metadata in the standardised XML format for uploading into a preservation metadata repository. The Java/XML tool comprises a generic application and a number of 'adapters' developed to extract the data from specific file types.[47] It has to be noted that the type of metadata that can be extracted is automatically limited to formal features that are part of the digital file, such as the number of pixels, the date of creation and the version of the data format. Metadata that expresses the semantics and syntax of the digital object cannot be extracted automatically.

Metadata supports the durability of digital objects and facilitates curation activities. A wide range of metadata schemas exists and a number of designated communities are developing, using and maintaining these schemas. The interpretation of metadata elements that are part of a metadata schema can vary within different groups of users. Metadata elements are taken from existing metadata element sets and adapted for local use.[48] People

tend to mix and match terms from multiple standards in order to meet the descriptive needs of a particular project or service. The set of metadata elements that are drawn from a number of metadata schemas and optimised for a particular local application is called an 'application profile'. Application profiles reuse existing metadata elements. The creation of metadata registries facilitates the easy mixing and matching of metadata elements. A metadata registry stimulates the realisation of an efficient method for the creation, dissemination and application of metadata elements of digital objects and can be considered as a relevant tool for data curation activities.[49]

5.5.4 Data emulation and data migration services

The two main ways to overcome technological obsolescence of digital objects are data emulation and data migration. Emulation is the process of imitating obsolete systems on future generations of computers. Migration implies the re-encoding in new formats before the old format becomes obsolete.

An example of an emulation tool is the digital asset preservation tool of IBM.[50] This tool is a demonstration of the UVC (Universal Virtual Computer) solution. The basic idea of the UVC method is that the bitstream representing the data object is stored together with a logical view of the data. A logical view of the data is easy to understand because it follows the way the user normally thinks about the data, rather than the internal representation often designed for efficiency. Not only the logical view but also the specification for processing the data on a future platform is archived. The processing specification is based on a Universal Virtual Computer. The UVC programme is independent of the architecture of the computer on which it runs. A UVC interpreter has to be written for any future target machine.

The Typed Object Model (TOM) is an example of an approach that applies the migration strategy.[51] TOM is a system for managing diverse data formats. It is a data model for describing a wide variety of data types and formats. The model makes it possible to: 1) explain what a given data format is, 2) interpret the format to extract information from the data, and 3) convert or migrate the data into more usable formats.

5.5.5 Persistent identification of data objects

Persistent identification of digital objects is an important component of a data curation infrastructure. Requirements for the persistence of the identifiers are their authority, reliability and functionality throughout the life cycle of the digital object. A persistent identifier tracks a specific object regardless of its physical location or current ownership.

The application of persistent identifiers for digital data objects consists of five steps: 1) selection of a persistent identifier scheme, 2) establishment of a naming authority, 3) creation of persistent identifiers according to the

identifier scheme chosen in step 1, 4) registration of the persistent identifiers. The identifiers must be translated to locations. For this a resolution service is required, 5) usage of the persistent identifiers. Six systems can be used for the persistent identification of digital data objects: 1) Universal Resource Name (URN), 2) the 'info' URI, 3) the Persistent Universal Resource Locator (PURL), 4) the 'Handle system', 5) the 'Digital Object Identifier' (DOI), and 6) the 'Archival Resource Key' (ARK).[52]

All available digital identifier policies and all systems for the creation and application of persistent identifiers require registration and resolutions services. Successful implementation requires institutional support and management. In principle, the only guarantee of the usefulness and persistence of identifier systems is the commitment of the organisations that assign, manage and resolve identifiers.

5.5.6 Trusted digital repositories

Increasingly the concept trusted digital repository is used in relation to digital preservation and data curation. The creation of added value is enabled by a high-quality data storage infrastructure. A clear and shared definition of this concept does not exist. The Nestor Working Group defines a digital repository as an organisation that has assumed responsibility for the long-term preservation and long-term accessibility of digital objects, ensuring their usability by a specified target group.[53] Trustworthiness is the capacity of a system to operate in accordance with its objectives and specifications. From an IT security perspective, the fundamental considerations are integrity, authenticity, confidentiality and availability. IT security is therefore an important prerequisite for trusted digital repositories.[54]

The *Nestor Criteria Catalogue for Trusted Digital Repositories* makes a distinction between criteria related to the organisational framework, criteria concerning the management of the digital object and criteria concerning the infrastructure and security. The *Nestor Criteria Catalogue* has received widespread international recognition since its publication in 2006. For instance, it now serves as the basis for various international activities such as the involvement of Nestor in international standardisation via ISO.

The DRAMBORA toolkit (Digital Repository Audit Method Based on Risk Assessment) is available to facilitate internal audit by providing repository administrators with a means to assess their capabilities, identify their weaknesses and recognise their strengths.[55]

5.5.7 Trustworthy digital objects

A trusted digital repository is aimed at long-term preservation and long-term access of digital objects. The focus is on the organisational framework and on content management issues. One could also take the digital objects to be preserved as the starting point of view. Gladney introduces the 'TDO

methodology' for preserving anything that can be preserved, including 'dynamic' information.[56] TDO stands for 'trustworthy digital object'. These are digital objects that can speak to their own authenticity. They maintain a record of their change history so future users can know with certainty that the contents of the object are authentic.

Gladney explains the TDO methodology by emphasising the close analogy with long-established practices for preserving works on paper. This analogy between preserving works in digital form and works in paper form consists of five parts: 1) replication of the information carrier in multiple independent repositories makes the information durable; 2) inventories and catalogues help consumers to find any preserved document; 3) augmenting a source version with representations in the appropriate lingua franca ensures that consumers can use any preserved document; 4) a document is trustworthy by attaching evidence, which might include signatures, embedded in a socially acceptable infrastructure; 5) information technology complexity is hidden from end users by a combination of education and of refined design.[57]

A number of the components that Gladney uses to construct a TDO are already mentioned in this report, such as persistent identifiers, emulation services, the XML data format and format repositories.[58] Specific to Gladney's approach is the application of knowledge theory principles and scientific philosophy to determine the essential digital preservation features of digital objects. Concerning the communication of encoded information (a process relevant for long-term access to digital data) the intended information must be distinguished from accidental, ephemeral information related to the objects.

5.5.8 Interoperability

Storage of digital objects in distributed, interoperable repositories is gaining ground as an efficient model to provide permanent access to digital objects where creators have control over the integrity of their work and the right to be properly acknowledged. In the scientific community the open access model has been implemented in several cases. A protocol for metadata harvesting, developed by the Open Archives Initiative (OAI-PMH), is an important construct for the implementation of permanent access and thus durable storage.[59] The OAI-PMH protocol permits metadata harvesting.

The goal of the OAI-PMH is to supply and promote an application-independent interoperability framework that can be used by a variety of communities engaged in publishing content on the web. The OAI-PMH is a communication protocol or language with only six permitted verbs.[60] The protocol effectively removes the dependencies on system architecture and metadata compatibility.

The workshop 'Augmenting interoperability across scholarly repositories' in April 2006 discussed steps that could be taken to augment interoperabil-

ity across heterogeneous scholarly repositories. A data model was proposed that intends to provide a common representation of digital objects in a set of heterogeneous digital repositories.[61]

5.5.9 Architectures for data curation

A number of tools and services exist that can be used to carry out data curation activities. These tools and services can be part of a system. The architecture of this system contains the structure of the tools and services and the relationships between them. Three systems architectures are discussed that contain components relevant for data curation. These architectures are the digital repository infrastructure as developed in by the DRIVER project the EASY architecture and the PANIC architecture.[62]

The aim of the DRIVER system is to allow existing repositories to deliver their content to larger communities of end-users through personalised services and interfaces. These repositories may contain the outcome of scientific research in any field, raw data, software, satellite pictures, tutorials, and multimedia and may conform to various models. The DRIVER repository infrastructure aims at collecting heterogeneous content and aggregating it to form a uniform 'information space', which delivers the original data sources through the same interpretation. The DRIVER system collects objects from different data sources. DRIVER-compliant repositories must publish their content through the OAI-PMH protocol (see section 5.8 of this report) and contain toll-free, accessible textual files. The DRIVER system provides end-users with uniform search interfaces to the heterogeneous content. The DRIVER architectural specification contains a detailed description of the DRIVER object model and the OAI-item repository model. The DRIVER object model was specifically devised to support a semantically and structurally uniform Information Space populated by 'objects collected from external heterogeneous repositories'.

The most important data curation task the DRIVER architecture will perform is the disclosure of digital research data available in a distributed environment. Obviously this will facilitate the re-use of existing data. The organisations that provide the DRIVER system with repositories are responsible for the durability of the research data.

Kramer describes another system architecture aimed at ingesting and publishing scientific datasets in the humanities and social sciences.[63] The system is developed and maintained by the Dutch scientific data archive DANS. Currently a number of components of the architecture are implemented and available on the web.[64] Researchers can use the EASY system to deposit and to retrieve scientific datasets. Data files can be uploaded through a web interface together with metadata that describes its content. The data that is deposited in the system has to be maintained and kept accessible for an indefinite period of time, since usability for future research of these data sets has an unknown expiration date. The infrastructure contains a dedicated data repository system called AIPstore.[65] This

component stores data without making any assumptions on the nature, format or contents of the data set or the metadata. The data storage is based on the open and durable XML data format.

Durability of the digital objects ingested in the AIPstore is an important design principle of the EASY archive system. The application of persistent identifiers (see section 5.5 of this report) is part of the system architecture. These two features are examples of data curation functions supported by the EASY system architecture.

A third example of system architecture possibly relevant for data curation is the integrated preservation framework PANIC.[66] The PANIC system is described as an integrated, extensible architecture based on preservation metadata, automatic notification services, software and format registries and semantic grid services. This offers a sustainable, dynamic approach to the long-term preservation of large collections of heterogeneous scientific data.[67] The PANIC system comprises of three main components: 1) preservation metadata generation tools, 2) obsolescence detection and notification services, and 3) preservation service description, discovery and invocation. The PANIC system uses a number of data curation tools and services mentioned in this report, such as the PRONOM and GDFR format registries (see section 5.1) and the UVC approach (see section 5.4) as components. The system integrates complementary preservation registries and services via semantic web services architecture. The web services (such as the format registries) are semantically described using a machine-processable ontology.

The tools and services assessed in this section of the report cover a diversity of functions related to the concept of data curation and illustrate that data curation can be implemented in practical situations. They will be further developed in the future and new tools and services will emerge.

5.6 Conclusion

Digital data curation as a concept became common in about the year 2000, but has now been recognised as a fundamental pillar of e-Science. Data curation and preservation of digital resources are seen as challenges that are difficult if not impossible for individual institutions to resolve on their own due to the complexity and scale involved. The curation of research data is no longer a simple side activity, but a significant element of research in its own right.[68]

Increasingly researchers are creating and using collections of digital research data. This requires digital data preservation and data curation activities. These are complex issues that require a strategic policy approach and development of an international infrastructure. Science is being transformed by accelerating change in information technology, with huge increases in computing power and network bandwidth, accompanied by an explosion in data volumes and information. The selection for preservation and curation requires good procedures for data and record management.

Information management policies are central to this and require co-ordination of policy at a high level.[69]

Digital research data has a great number of appearances. This is illustrated by the report 'Trusted digital repositories: attributes and responsibilities' (RLG/OCLC, 2002) that states that repositories contain 'heterogeneous collections held by cultural organisations'. The report concludes that research and the creation of tools to identify the significant attributes of digital materials that must be preserved is required. Since 2002 a great deal of work has been done resulting in many tools, systems and services that can be used to implement data curation activities. A number of these have been discussed in this chapter.

We are only just beginning to move towards the realisation of a robust, commonly agreed-upon data curation infrastructure. In 2005 a group of experts compiled a list of about 50 research issues in the field of digital curation and preservation.[70] This research agenda for the next decade contains many ambitious goals and a number of initiatives are on the way to reach these goals.

The aim of this chapter is to contribute to a better understanding of digital data curation issues and provide a number of concrete directions for the improvement of the research data infrastructure. Gladney optimistically writes: "if the technical and organisational challenges are overcome, digital preservation is likely to become a routine activity with priorities set by each institution's resource allocation process".[71] This situation has not yet been reached, but will hopefully be realised in the coming years.

6. Long-term preservation for institutional repositories

Barbara Sierman

6.1 Introduction

The growth of institutional repositories and their valuable digital content raises questions as to how to preserve this content for the long term. In this chapter the main topic is the long-term preservation of digital material and its consequences for institutional repositories.[1]

In 1999 Jeff Rothenberg was one of the first to raise the question on how we can preserve digital material over the years.[2] In his article he imagined that he left his grandchildren a CD-ROM and a letter, in which he told them that the way to his fortune could be found on the CD-ROM. But they found this CD-ROM in 2045. Could they find the treasure, could they read the obsolete CD-ROM? This article was the starting point of many discussions. During the last decade, many articles have been written and conferences organised around the theme of digital preservation. Everyday, new projects and insights are made public on the websites of the preservation communities. Despite all these efforts, digital preservation has not yet attained its full development. For institutional repositories, with their collections of science treasures waiting to be found by contemporary and future generations, the question of how to preserve this valuable material for the long term and how to keep this accessible over the years, becomes vital. Unfortunately, a clear and simple recipe with rules and guidelines on how to perform digital preservation in a consistent manner has not yet been written. Digital preservation is a relatively new area, and one in which a broad community is still searching for the best way to handle digital material for the long term. This means that starting points are still subject to discussion (like the OAIS model) and that, although the goal is clear, the necessary tools to reach this goal are still to be developed or to be improved.

In this chapter, one of the main goals is to raise awareness. It is important for every institutional repository manager to be aware of the aspects of digital preservation and to be familiar with the solutions and measures he/she is supposed to take for his/her repository.

This chapter will describe various aspects of digital preservation. Information about the object, such as file formats, will be discussed in paragraph 6.3, metadata about the object in relation to its environment in paragraph 6.5. The starting point of digital preservation – the theoretical OAIS model – will be discussed in paragraph 4, and preservation strategies in

paragraph 6.6. Finally, the organisational aspects of digital preservation are the subject of paragraph 6.7.

6.2 The rationale for digital preservation

6.2.1 What is digital preservation?

According to the standard description of the OAIS model (see paragraph 6.4), digital preservation is described as 'the act of maintaining information, in a correct and independently understandable form, over the long term'.[3] Many more definitions have followed in an attempt to translate this rather abstract description into a more practical one. Jones and Beagrie spoke of a 'series of managed activities necessary to ensure continued access to digital materials'.[4] This description reflects the actions to be taken: managing the data and the accessibility of these data. This doesn't just mean storing the bits and performing regular backups, but it involves extra effort needed to help understanding the data over the years. This is not a one-time action, but demands permanent attention and an organisation willing to take up this task.

6.2.2 Why should we pay attention to the digital preservation of IR material?

In 2003 the UNESCO adopted the *Charter on the Preservation of the Digital Heritage,* in which it brought into attention the digital preservation of cultural (digital) heritage.[5] Starting points were formulated and the member states signed their consequent responsibilities. A description of cultural heritage was given. Besides libraries, archives, museums and so on, institutional repositories are also part of the cultural heritage of a country, as a result of the content of their repository, the scientific output of a country. The number of institutional repositories is growing, both in Europe as well as in the rest of the world. The repositories have a valuable collection and a growing audience, consisting partly of the general public, but mainly of an academic audience, that is using the material in the repository. These consumers of the repository trust they will be able to have access to the repository over the years. This demand of the public requires that the repositories start to think about the measures to be taken to keep these repositories accessible for a long time. Nowadays the main focus of the institutional repositories is often on collecting material, storing it into the repositories, and making it accessible for a wide community of interested people. Digital preservation itself is often not yet part of the daily workflow. Nor is it clear why and how to perform digital preservation and who should take care of this process.

It is beyond discussion that institutional repositories have an important role in the digital preservation of their scientific output. After all, they are custodians of the scientific output from the very beginning. The moment when the paper, article or book is created (in this article the term 'digital object' will refer to all kinds of scientific output) is the moment when certain choices are made that are of great influence for the long-term preservation of this material. These choices concern the file format, the use of the parameters and the feasibilities the software offers, the metadata added to the document, and the medium in which the document is stored.

This chapter intends to show how repositories can influence the long-term preservation of their material.

6.2.3 What is so special about digital material?

Why is digital material so special? Digital material is created with special software, running on a dedicated technical environment, resulting in files with specific characteristics and behaviour. A combination of software and hardware is needed to create these digital objects. The resulting digital object is not readable by humans, but needs a similar environment of hard- and software to be read or viewed, or needs special readers or viewers.

This situation is not a stable one, as commercial enterprises create new versions of software and hardware, while older versions of the software and hardware become obsolete.

Each year new generations of hardware components are brought onto the market. For example: floppy disks developed from the 5.25-inch floppies to 3.5-inch floppies, and nowadays new computers are not being sold with floppy drives anymore. Hardware degenerates by the use of plastics, rubber, etc. Or it becomes obsolete, simply because its functionality is no longer maintained. The same goes for software, where new versions do not always support all facilities of older versions of software (compatibility) and where some software completely disappears (WordStar, for example) or is no longer supported after a certain period of time.

The storage of digital material on hard disk, tape or another medium is also a task that requires constant monitoring of the carriers. The fragility of magnetic and optical carriers and their physical deterioration over the years must be carefully watched. The constantly changing environment of the digital object, be it software or hardware, and long-term storage are the main ingredients that form a threat to the long-term preservation of digital material. Hence, special actions need to be taken. Digital preservation is about these special actions. These actions need to be performed during the whole life cycle of the object. So performing digital preservation in a professional way requires a permanent commitment of a solid organisation. The institution that preserves material for future generations holds a long-term commitment and needs to organise these long-term obligations; through funding, research and by creating an environment of people and material that is capable of dealing with new challenges. How this relates to the tasks

and functions of the institutional repository will be discussed in paragraph 6.7.

6.2.4 Is there a problem?

Are there any problems for the institutional repositories with regard to the long-term preservation of, and access to, their collections? At least one aspect is relatively new in the world of digital repositories, and of great influence. In the paper-based world, an article was published in a journal, of which thousands of paper copies were distributed over the world. Libraries with a subscription to the journal took care of preserving these journals. But e-prints in the repository often have no paper counterpart anymore. This digital object is the only manifestation of the article (although its digital character gives the opportunity to multiply it endlessly). Who will take care of this digital paper and preserve this? Is this the function of the institutional repository?

The sceptics among us might argue that the problem of digital preservation of these publications in the repositories is exaggerated. There are examples of repositories like arXiv, which exist since the early 1990s and all information is still readable. Is digital preservation an invention like the millennium bug? Recovering damaged disks is a highly developed trade. Specialised suppliers are able to recover hard disks that have been placed under water, or even have been in sea water for years, be it at huge costs.

The example often used to show the need for digital preservation is the Domesday book, a BBC project to celebrate the 900th birthday of the Domesday book by asking people all over the UK to help and create a new contemporary version of the book from 1086. The project results were stored on highly advanced media. After a few years, the personal contributions of hundreds of British citizens were no longer accessible. With a lot of efforts and costs, eventually the members of the CAMiLEON (Creative Archiving at Michigan and Leeds: Emulating the Old on the New) project succeeded in making it accessible again.[6]

A project group at Cornell University did a study among its staff and students, asking them if they had an old disk with unreadable content. One of the results of this experiment showed that it was not often the carriers that caused the problem (they were mainly well looked after), but the obsolescence of the software, the lack of information, documentation and handbooks of the used and outdated software.[7]

This lack of information about the original object and context seems to be the challenge for long-term preservation, since this information is lost if it is not explicitly kept for the long term.

Another challenge is the character of the digital object itself. As technology evolves, digital objects are growing in complexity. A few years ago, digital objects were created in a file format that was relatively straightforward regarding the parameters and features that could be chosen. For example, an article was mainly a text file, with some illustrations. Nowadays, digital

objects may consist of more complex file formats, allowing a combination of sound, movie and text in one digital object. Sometimes even databases are added. To enable future use of these file formats, and to do right to the features of these file formats with their interrelations, requires thorough knowledge of the digital object. As the first custodians in the life cycle of the digital objects, the management of the institutional repositories should be aware of this.

6.2.5 Two goals of digital preservation: storage and access

To get an idea of the implications of digital preservation, as 'a series of managed activities necessary to ensure continued access to digital materials', it is important to make a distinction between the two main goals of digital preservation: storage of the digital material and permanent access to this material.

Storage

Over the years the storage problem itself has had a lot of attention from the producers of storage systems. Hardware has undergone major improvements, the storage capacity has grown immensely (petabytes and terabytes, while a few years ago gigabyte was the maximum), and the storage media are more diverse; the latest invention is a holographic disk which can contain even more information. Intelligent storage systems take care of the monitoring of the stored data and give warnings when things go wrong. On this technical part there is lot of experience.

Before storing the material, a decision has to be made on to how to store the material, for example in which file format. This question has no straightforward answer yet and will be discussed in paragraph 6.3.

Long-term access

Digital preservation is about permanent access: to enable the future user to use the digital object in the appropriate way. This point is less straightforward than storage of bits. It requires an organisation to develop a preservation policy in which the adequate use of preservation strategies like normalisation and migration are planned. Appropriate use of these strategies asks for information about the object. This information, the metadata, will offer the future user a responsible rendering of the object. As this future user will operate in an environment that at this moment is totally unknown, we need to take action from the very start of the digital object to enable future rendering. Some aim at offering the future user 'the original look and feel' of the digital object. Others will leave it up to next generations to create an environment for rendering the object and leave the details to them.

6.2.6 Who should take the responsibility?

Institutional repositories collect the material of their institution and offer this scientific material to a wide range of people. Should the organisation behind the institutional repository be the one to fulfil long-term preservation obligations? Some argue that digital preservation is not a problem for institutional repositories.[8] Their repository contains e-prints or preprints of articles. Publishers will publish these preprints officially in their digital or paper journals. In many countries the national library takes care of these publications as part of their depository task and there is no need for an institutional repository to do the job again. This way of thinking is valid, as long as the repository only contains e-prints of publications. But often a repository contains more information, like reports, learning material, databases and research data, and all kinds of scientific output that will not be published officially. The users of the institutional repository will expect to have access to this information for a long time. All this information will be lost if special care is not taken. This special care means digital preservation in one way or another. So the management of institutional repository needs to think about its role and develop a policy describing whether or not the institutional repository sees digital preservation as a responsibility. Or their business model might be founded on a collaborative approach, as is done in the Dutch SURF-DARE initiative.[9]

So there are many reasons why the management of a repository should think about long-term preservation: repositories contain more information than the official publications only, and their audience expects them to keep this information accessible (as was discussed in section 6.2.2). Another reason is that the repository contains many 'born digital' objects, and as such, is in the position to gather a lot of necessary information about the object, needed for long-term preservation.

6.2.7 Current research

Over the years, a lot of attention has been paid to various aspects of storing digital material and the design of repositories. Several initiatives such as DSpace[10] and Fedora[11] offered organisations a place to store their digital material, while libraries and archives designed and implemented digital repositories or e-depots. Current research focuses on the development of tools needed to characterise objects, perform preservation planning, and so on. Several European projects, each with different goals, will present their results in the next few years.[12] Although the focus of these projects is not restricted to institutional repositories, their results will be of benefit for every organisation performing digital preservation.

– DPE (Digital Preservation Europe) is a project that aims to foster collaboration and synergy between existing national initiatives across the European Research Area. Besides from several universities and archives,

the national libraries of Denmark and the Czech Republic are partners in this project.[13]
- Caspar (Cultural, Artistic, and Scientific Knowledge for Preservation, Access and Retrieval, 2006-2009) will build a preservation framework for heterogeneous data, along with a variety of innovative applications. They focus on scientific data, as well as on cultural data and contemporary arts.[14]
- Planets (Digital Preservation Research and Technology, 2006-2010) is a project consisting of 16 partners, including the National libraries of Austria, Denmark, the Netherlands and the UK, together with archives and universities in Vienna, Glasgow and Cologne as well as commercial partners.[15] In this project the goals are concentrated on 'preservation planning': which planning instruments does an institution with a digital collection need, given the organisation's policy, budget, content and intended use of this content. This approach will result in a 'decision support system' to assist the preservation institutions in making a right decision. To support this decision-making process, a wide range of tools will be developed to actually execute a chosen plan Further development will be done on preservation strategies like migration and emulation. A test bed will be created where the participants of the project can test their results in a representative environment. Other partners (commercial or not) may test their digital preservation products here too. The deliverables of the project will be mostly open source.

6.3 Digital material

6.3.1 Digitised and 'born digital' material

An institutional repository will contain different kinds of material: publications in various stages like preprints, post-prints, but also learning materials etc. The majority of these documents will be produced locally, at the institution and start off as a digital document; these are the so-called 'born digital' objects. In contrast with these objects are the digitised objects: they started their life as an analogue object and digitisation gave them an extra digital life. Is this distinction between born digital and digitised important? Yes, for digital preservation reasons it is. In the case of digitised objects, there is, at least in theory, but most of the time in practice as well, an analogue original file, which is the starting point of preservation. The characteristics of this original file have to be included in digital preservation decisions. Apart from that, the analogue original might help in case the digital one gets lost, or in case it is determined that the quality of the digital object was insufficient.

In contrast with the digitised objects, the born digital objects have no predecessor, so the born digital object should contain all the information

there is about its characteristics and properties. Once created, the original result can be copied, but cannot be recreated.

6.3.2 Creating a digital object

Either born digital or digitised, every moment a digital object is created, certain choices are made which can influence the possibilities for long-term preservation. One of the important choices that is made is the choice of the file format. There are hundreds of file formats and not all formats are suitable for long-term preservation. Software programs, which ultimately create the file format, offer the user a range of options that might influence the preservation of the object. The creator or author might not be aware of these risks. For example, he chooses the PDF format and sets a password on the file, because he does not want somebody else to copy and paste from his text. But over the years, no one will know the password any more. This might not be a problem for accessibility, because the file is still readable. But this choice might hinder certain preservation actions with the object, like migrating it to another format, in case the PDF format becomes obsolete.

The digital repository manager should seek a balance between easy depositing and costly preservation operations. The file format choice can be influenced to a certain level, by limiting the file formats allowed (to open standard-based file formats, whenever possible) or by performing normalisation on the digital objects. Guidelines will add to better understanding and raising awareness of the authors.[16] Although there is a tension between limiting file formats allowed in the repository and the freedom of the author to use his favourite programme, a repository can help the author to show him the risks of his choices. On the internet several guidelines can be found in this area.

6.3.3 Preservation levels

Not every institution has the means and (technical) opportunities to guarantee long-term preservation of its digital objects. It might be that an institution starts with a digital collection, preserves this at a minimal level, and after a certain period hands the content of the repository over to an organisation that is well equipped to perform digital preservation. A 'preservation level' enables the institution to show the user of the object to which level the repository was able to take its responsibility; it shows to which level the institution has preserved the objects in the repository. In this way, an institution can await the moment technology offers new opportunities to treat the object correctly according to a higher preservation level. Although there is not a fixed list of preservation levels, there are some widely accepted levels. The basic preservation level is 'bit stream preservation'; raw data are being stored and kept exactly as they have been delivered. Although this

requires a well-qualified IT environment, the future user will not have the guarantee that the object is rendered as it was originally, because the object lacks information about the interpretation of the bit stream. Other flavours of preservation levels are for example 'access preservation' or 'representation preservation'.

6.3.4 Recognising preservation risks

Over the years the possibilities of creating digital documents have grown immensely. An MS Word file is not only plain text, but may contain several different fonts, images, links to spreadsheets and links to URLs. Once a PDF file is created from this file, the links are fixed, and the font set can be embedded while creating the PDF file. The amount of different file formats is growing. From preservation point of view, the greater the diversity of file formats in the repository, the greater the risk and task of preserving this material.[17] There is always a certain tension between choosing the best file format for preservation and the freedom of the creator of the object. Restrictions should not lead to authors not depositing their material. Normalisation (migrating to a more preservation fitted file format) might be a solution.

Some criteria can be distinguished for file formats to be best for preservation reasons. Jones formulated the following:[18]

- Is the file format an open standard/format?
  If the file format is an open standard/format, then the specifications of the file format are publicly available. This offers an opportunity to create ways of accessing the files.
- Is the file format widely used?
  The assumption is, that if a file format is widely used, there will always be solutions to keep the documents accessible, as it is in the interest of many people.
- Is the file format and associated technology likely to be preserved?
- Is the content of the file human readable?
  This is an argument for example for XML, which is human readable, although one also needs the translation to the meaning of the different entities
- Is the file format itself human readable?

6.3.5 Preferred file formats

Industrial developers recognise more and more the hesitations of governmental organisations and archives towards proprietary file formats. Adobe recently made the fist step to get its PDF 1.7 format certified as an ISO standard. The Open Document Format (SUN/IBM) is already an ISO standard and Microsoft is working to get Office Open xml standardised. Another choice is translating files to XML; this however requires storage of

documentation with the file (style sheets, etc.) and is not fit for every file format. It is always important to store the original file as well, as files that cannot be rendered today, might be in the future, for example with the help of emulation (see 6.6.4).

6.3.6 Determining file formats

For long-term preservation of digital objects it is crucial to know all characteristics of the objects that are stored in the repository. One of the characteristics of the object is the file format and version. How can a repository owner see which file formats are used and whether the author did use certain features or not? Files have file extensions, like .pdf or .txt. But these file extensions are not unique, a .doc document can be created with several different word processor programs and the extension can be changed by the author himself. So, the extension itself is not trustworthy and unique enough, it does not provide any information on the characteristics of the object. Above all, it does not give any information on the version of the file format. A better way to retrieve the information on the file format is to extract it from the file itself, where this information is also present. Specific programs are able to extract this file format information before archiving the objects. The extracted information can then be stored safely as metadata with the object.

One of the metadata extraction programs is JSTOR/Harvard Object Validation Environment (JHOVE), an open source tool developed at JSTOR and Harvard University Library, that can identify, validate and characterise the file format for a limited set of widely used formats. JHOVE has three main functions: identification, validation and characterisation.

- With *identification* the question is answered: which file format is it? The answer is based on internal information in the file (and not just the file extension).
- With *validation* the question is, does the object meet the requirements of the file format? The three criteria are:
  - Is the file format syntactically correct? If so, it is well-formed.
  - Is the file syntactically and semantically correct? If so, the file is valid.
  - Is the representation information correct? If so, the file is consistent.
- Last but not least, JHOVE can perform format *characterisation* and identify format-specific characteristics, for example whether embedded fonts are used and whether these embedded fonts are enclosed in the document. Fonts are a special problem in preserving the original look and feel: if the fonts are not embedded within the digital object, rendering on a different computer where those fonts are not installed might lead to another rendering of the object, and sometimes to errors in representation of text, formulas and tables.

It is important to notice that the use of JHOVE will not affect the original object; the programme just extracts technical metadata. The drawbacks of

JHOVE are that there is at this moment only a module available for a limited set of 12 file formats; it produces a large amount of unqualified technical metadata; there is only limited documentation on the module, and it is not suitable for complex file formats.

Other programs exist that have some of the functionalities JHOVE has. For example DROID, developed by the National Archives in the UK, is a programme that can be used for file format identification.[19] DROID offers only the identification function, but is able to do this for a wider range of file formats and as such is a suitable tool.

6.3.7 File format registries

Some information is closely related to the digital object and should be stored together with the object that will be preserved, like information about the use of character sets. Other information about file formats is more general, like information about the supplier of the software, the maintenance of the software, the environment needed to run the software, the necessary tools to render the software, newer versions or perhaps even the successor. It is not necessary for a repository to store this general information on its own. Worldwide several initiatives have started to store this information in so called file format registries, to be consulted by everyone, as they are freely available on the internet.

One of the examples is PRONOM, an initiative of the National Archives of the UK, describing the file format information of over 130 file formats.[20] The PRONOM registry is the first initiative to come to a registry of file formats with a unique identifier for file/version formats, the 'PRONOM Unique Identifier' (PUID). The idea is to have a worldwide registry of file formats, making it possible for repository owners to add this number to the technical metadata of the object and link the objects to the (centrally located) file format information.

Another initiative is the Global Digital Format Registry, supported by the NDIIPP project (a national project in the United States on digital preservation).[21] It is to be expected that these initiatives will lead to a worldwide system of unique identifiers, where the PUID will be leading.

However, the development is that file types are getting more complicated, and that file formats can offer more possibilities. The newborn 'container formats', like PDF and JPEG2000, can contain different kinds of data. These container formats are a big challenge for digital preservation due to their complexity. To extract file format information requires the repository system to allow checking as part of the storage process (in OAIS terminology called the 'ingest flow'). This ingest flow should also allow the use of different checking programs for different purposes.

6.3.8 Persistent identifiers

When storing a digital document in the repository, it is important to identify this object uniquely, for now and for the long run. Even if the repository moves to another archive, the objects should keep a unique identifier, the 'persistent identifier'. This identifier enables the researchers and repository managers to identify the object, and use it in the scholarly process. If a national deposit library stores an e-book, this is the final version of the book from the publisher. In the academic world however, the institutional repository will have to deal with several versions of the original document. Future users who want to access the document will also see the history of the document (like different versions and the relation between them), which is called the provenance information. They want to see the difference between the versions of an object, and know exactly whether they have the latest (peer-reviewed) version or a previous one. These relationships should be reflected in the database of the institutional repository.

The only way to deal with this properly is to have a unique and persistent relationship between the different versions. This can be realised by adding a unique identifier to the digital object and using this as a reference. Although several systems have been developed, like URN, the Handle system, the DOI and the PRUL, the use of persistent identifiers is not yet widespread.[22]

6.4 OAIS – Open Archival Information System

6.4.1 Background

During the 1990s the Consultative Committee for Space Data Systems (CCSDS) published 'Recommendations on preserving material'. These recommendations were soon adopted in different environments involved in digital preservation. The model which was laid down in these recommendations, the 'Reference Model for an Open Archival Information System' (OAIS), became an ISO standard in 2003.[23] The OAIS model is a reference model. This means that it is not a guideline on how to implement a digital preservation system, but it describes an overall concept of the functions and activities, related to digital preservation and it states which responsibilities a repository should fulfil to be OAIS compliant. Although the text of the OAIS model is not always explicit about how an archive should fulfil the requirements mentioned, different initiatives have made a translation from the OAIS model to the practical requirements an institution should meet. More about this will be explained in 6.7.1, where the audit of repositories is the main subject.

Apart from the fact that this model is a thorough and robust one, it also offers a basic set of terminology for everyone involved in digital preservation, and it is widely used in this way. The OAIS model is not perfect or

final. The frequent discussions in the digital preservation community have led to new insights, technology has evolved, and more people have experience in implementing a repository, based on the ideas of OAIS. As it is an ISO standard, regular (five-year) reviews will be held, which offers the community a chance to improve the model. This year (2007/2008) the model will be reviewed for the first time.

The OAIS model defines a set of responsibilities for an archive that wants to operate as an OAIS-compliant archive. In OAIS an archive is described as 'an organisation that intends to preserve information for access and use by a designated community'. In this sense an institutional repository is an archive, as is a data centre.

This section describes the minimum set of mandatory responsibilities an archive has to follow to function as an OAIS compliant archive. First, this section gives an overview of the main topics that form the OAIS model, the actors involved in the process (6.4.2), the concept of the functional model (6.4.3), and the concept of the digital object.

6.4.2 The OAIS model

The actors

Three actors are distinguished in the model: the producer, the consumer and the management of the repository.

Producer

This is the role played by those persons or client systems which provide the information to be preserved. This can include other OAIS or internal OAIS persons or systems. There is a submission agreement between the OAIS and the producer, in which various items are described, like the layout of the SIPS (see later), the submission times, etc.[24]

Consumer

The consumer of the archive is the ultimate client for whom all the effort is done. The consumer might be a contemporary user or some future user who wants to have access to a digital object from the past. The archive had this user in mind when making decisions about access methods and discussing ways to render the digital objects in the future. The user is part of what OAIS calls the 'designated community': 'an identified group of potential consumers who should be able to understand a particular set of information'.[25] For this user, however, sometimes it will not be enough to save only the digital object, as he might need extra information to fully understand the digital object. Apart from digital preservation, an archive also takes care of information preservation. There is as yet no tradition or fixed method for how to fulfil this task, and there is no consensus about the necessary extra documentation needed to do information preservation. In practice, this notion of designated community is not always easy to translate to the situation of the repository, as a lot of them will have a large group

of consumers among their audience. For example, a national archive might have 'all English speaking people' in their community. However, if a repository archives highly specialised material, used by a relatively small group of specialists, it requires extra attention to the measures to be taken to keep this material understandable by the intended user over the long term.

Management
Management does not refer to the staff that monitors the day-to-day operations, but to the management of the organisation operating the repository. This management is involved in formulating and fulfilling the charter and scope of the OAIS. In the audit of the repositories their responsibilities and tasks are more differentiated.

6.4.3 Functions in the OAIS archive[26]

The digital object, handed over by the producer to the management of the organisation and later retrieved by the consumer, undergoes a process of actions in the repository, displayed in figure 12 below.

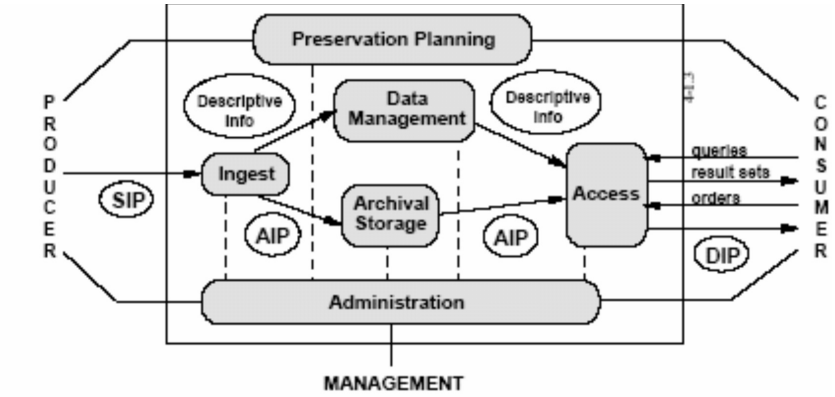

Figure 12 – OAIS functional entities[27]

The functional model consists of six blocks: Ingest, Access, Data Management, Archival Storage, Administration and Preservation Planning. A seventh block, that is always present but not part of the figure, is Common Services.

Ingest
The Ingest function provides the services and functions to cover several activities around the Submission Information Package (the digital object), ultimately resulting in an Archival Information Package (AIP) that is fully prepared to be stored. This means, among other things, that quality assurance checks have been performed and that the necessary descriptive infor-

mation has been generated to make it possible for users to find the digital object in the archive.

Archival Storage
This entity provides the services and functions for the storage, maintenance and retrieval of AIPs. Functions like receiving Archival Information Packages from Ingest and adding them to permanent storage, managing the storage hierarchy, refreshing the media on which the holdings are stored, perform routine and special error checking, provide disaster recovery capabilities and providing AIPs to Access to fulfil orders.

Data Management
Provides the services and functions for populating, maintaining and accessing both Descriptive Information like descriptive metadata (which facilitates the access to the object) and administrative metadata used to manage the archive. This includes performing queries on the data, providing reports of these queries, performing database updates and administering the archive database functions.

Administration
Provides the services and functions for the overall operation of the archive system.

Access
Provides the services and functions that support Consumers in determining the existence, description, location and availability of information stored in the OAIS, and allowing Consumers to request and receive information products.

Preservation Planning
Provides the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains accessible to the Designated User Community over the long term, even if the original computing environment becomes obsolete.

Common Services
Needs no extra description in the OAIS model, as this function is so pervasive, but refers to issues like security, network and operating system services.

The information package
As the functions and actors of the OAIS have passed in review, the attention should now go to the subject of the archive: the digital object, or in the OAIS terminology, the 'Information Package', see figure 13. The "Information Package is a conceptual container of two types of information, called Content Information", which is the actual digital object that the repository

wants to preserve, and the "Preservation Description Information (PDI) The Content Information and PDI are viewed as being encapsulated and identifiable by the Packaging Information. The resulting package is viewed as being discoverable by virtue of the Descriptive Information.

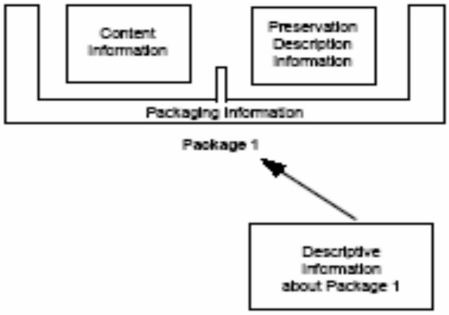

Figure 13 – Information package concepts and relationships[28]

"The Content Information is that information which is the original target of preservation, the digital object." Yet, it contains more: added to the digital object is the information needed to make the object understandable to the Designated Community, the so called "Representation Information", which is all information needed for rendering, like which soft- and hardware environment is needed.

Closely connected to this Content Information is the Preservation Description Information (PDI), in which all information that is needed to preserve the content information is stored, plus the information about the environment in which the Content information (the original digital object plus the extra information needed to understand the object correctly) was created. Different types of information are capsulated in this PDI:

- "Provenance describes the source of the Content Information, who has had custody of it since its origination, and its history (including processing history).
- Context describes how the Content Information relates to other information outside the Information Package. For example, it would describe why the Content Information was produced, and it may include a description of how it relates to another Content Information object that is available.
- Reference provides one or more identifiers, or systems of identifiers, by which the Content Information may be uniquely identified. Examples include an ISBN number for a book, or a set of attributes that distinguish one instance of Content Information from another.
- Fixity provides a wrapper, or protective shield, that protects the Content Information from undocumented alteration. For example, it may involve a check sum over the Content Information of a digital Information Package.

The Packaging Information is that information, which, either actually or logically, binds, identifies and relates the Content Information and PDI. (...)The Descriptive Information is that information which is used to discover which package has the Content Information of interest."[29]

This described information package is the subject of the OAIS archive. But in the model of OAIS, the information package gets different names, depending on its role in the functional entities of the archive. As shown in figure 12, the Information Package is called SIP (Submission Information Package) when the Information Package, coming from the producer, is sent to the OAIS archive. Once submitted and archived, it is called an Archival Information Package (AIP). When a member of the Designated Community via descriptive information retrieves one AIP or a set of related AIPs, he will see the Dissemination Information package (DIP). In all cases the information package consists of the digital object but the information of the PDI is also connected.

6.4.4 Mandatory responsibilities for an archive according to OAIS

The OAIS model defines a set of responsibilities for an archive (be it a repository or a data centre), that wants to operate as an OAIS archive. The following set is a minimal set of mandatory responsibilities:[30]

- "Negotiate for and accept appropriate information from information Producers" The organisation will define the criteria that will help in determining the types of information that it is willing to accept. This includes also ideas about subjects, information sources, techniques used to represent the information (format, media) and the level of completeness of the object (not only the object, but also the extra necessary metadata to understand the object correctly). This will be discussed in 6.5.
- "Obtain sufficient control of the information produced to the level needed to ensure Long-Term Preservation." It is likely that during the lifecycle of the object in the repository, preservation strategies require that a migration of the object is undertaken, which will change the digital object. Adding metadata, as a result of new insights, might also ask for enhancing the object information. The repository must have obtained the rights to perform this and other strategies, not being hindered by (intellectual) property rights etc. On the other hand, the repository cannot act in contrast with legal restrictions and should take the appropriate measures to honour these legal requirements.
- "Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided", as this highly influences the metadata to be stored. The Designated Community consists of the expected consumers of the repository, now and in the future. The repository should do its best to take their wishes into

account, although it is obvious that the consumers of a repository will change over the years.
- "Ensure that the information to be preserved is **independently understandable** to the Designated Community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information." In the case of institutional repositories this is an important responsibility. Due to the character of their digital objects, often intended for a specialised group of insiders, special care should be taken to meet this responsibility. In practice it is not quite clear how to do this. Should the repository also store e-books, explaining the terminology used in a certain group of scientists during a certain period?
- "Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original."
- "Make the preserved information available to the Designated Community." This includes that the user of the information will get the necessary information to find the objects (access information), but also that the user can be confident that he sees the intended object as it was when it was stored (authenticity).

6.4.5 The OAIS model for institutional repositories

The question remains whether the OAIS model is applicable for institutional repositories, since repositories are functioning as an archive. In the preface to the OAIS model, it is stated that the model is applicable for 'any archive'. In this definition, institutional repositories are included, as the repositories are archives with the intention to secure the scientific output for future generations, although the current institutional repositories might not be the ones who will carry out this task over the years. The conceptual character of the OAIS model allows institutional repositories to be compliant without much extra effort, as was the conclusion of a study into the relation between institutional repositories and OAIS. The use of the OAIS model will help to ensure good practice.[31] A study in which the National Archives and the UK Data Centre did some self assessment and investigated whether they were OAIS compliant resulted in a similar conclusion: 'Insofar as the OAIS model is only for reference and is designed for all types of archives, it becomes clear that any institution with a responsibility for preservation could meet these high level requirements'.[32] How far the requirements are met by an archive or repository might become clear after they have been audited. In 6.7.1 the standards for trusted repositories will be discussed, and the OAIS model will be translated into more measurable practices. This might help the institutional repositories to embed preservation in their workflows and translate their responsibilities into daily practice.

However, both reports concluded, and this is a widely accepted view, that OAIS is a valuable concept and offers the digital preservation community a shared set of concepts and terminology.

6.5 Metadata

Metadata is defined as 'data about data'. For digital objects, metadata is needed to support their functionality over the years, as the object itself does not give enough information. When trying to read an eighteenth-century book or medieval incunabula, the curator might give you instructions like wearing gloves, using a penknife to cut the still uncut pages or to use a supporting lectern and magnifier to study details. However this advice is not essential for reading the book, which is accessible without these tools. Digital objects cannot do without instructions on how to use them properly, as without this information the object is just a bit stream. A book without an index or cover is still readable, but a digital document without information about its structure, file format and bibliographic entries is hard to handle. Information about the environment and the interpreter to render the bit stream into meaningful information for the user is essential. And the best moment to configure this kind of information is at the moment of creation or soon after, when the details of the environment that is used are still available.

Under the umbrella of the general term 'metadata', this information is categorised. Metadata is of vital importance to undertake proper digital preservation. Carefully collected metadata enables the future rendering of the object, creates trustworthiness as it will prove the authenticity of the object, and helps the future user to identify the object over the years. To meet these requirements, several metadata sets have been developed and standardised.

6.5.1 Types of metadata

Metadata come in all sorts and varieties: descriptive metadata, administrative metadata, technical metadata, rights metadata, preservation metadata, etc. Digital objects need more than just descriptive metadata, which describe how to identify the object, but don't say anything on how to render the object, the software/hardware that is needed, etc.

There is not always a clear demarcation between the various sorts of metadata, which might cause some confusion, as you can find structural information under the heading of preservation metadata, for example. It is more practical to have the function of metadata in mind then to strictly try to classify them in different sorts. Keeping the main goals of digital preservation in mind, permanent storage and access, the following types of metadata can be distinguished;

– to identify the object: descriptive metadata like bibliographic information
– to preserve the object over time: preservation metadata
– to take care of long term access: structural metadata, rights and representation metadata.

6.5.2 Descriptive metadata

Although there are several standards for describing the bibliographic metadata, for institutional repositories, the Dublin Core set of metadata is almost the de facto standard. The main reason for this is that, apart from being a compact and flexible set of bibliographic items, it is an accepted standard in the OAI-PMH protocol, which is especially designed for the exchange of information between repositories.

The Dublin Core Metadata Initiative aims to develop an extensible set of basic metadata elements that describes a document for digital preservation.[33] The Dublin Core Standard is widely used and consists of 15 elements, all of which may be used optionally and as often as one deems necessary. While entering the required information, it is advisable to use as much standards as possible, for example the ISO standard for dates and language.

6.5.3 Structural metadata

As the digital objects become more and more complex, no longer consisting of one file, but more and more a collection of several interrelated files, structural metadata are needed to explain this structure to future users. Metadata schemas that support this information and offer the opportunity to describe one coherent set, are for example METS and MPEG 21-DIDL. These container schemas are able to combine several metadata schemas into one overall description, with nesting into several layers and mixing bibliographic, structural and preservation metadata into one set. By using METS or MPEG 21-DIDL the repository is free to choose which standard for bibliographic metadata to use.

6.5.4 Preservation metadata

After several years of study and interim reports, in 2005 the *Premis Data Dictionary* was published, in which an extensive set of preservation metadata is described.[34] The aim was especially to define a set of metadata that would be applicable in the digital preservation community. The great value of the resulting *Premis Data Dictionary* is that it helps organisations to determine which elements they need to gather for preservation purposes.

Also, they can use the dictionary as a checklist and so raise awareness in the organisation about collecting preservation metadata.

The dictionary consists of more than 120 items. There is a limited set of mandatory elements. The experience of repositories that implement Premis and evaluate the elements, will eventually lead to a list of recommendations. The repository manager is helped by the information in the dictionary that these elements cover 'the information a repository uses to support the digital preservation process'.[35] It nevertheless remains the repository manager's decision, which elements he will collect and where he stores this information. This can be done together with the object. Yet certain information, for example about rights, might be stored in a safe place at the university, where all the information regarding contracts and licenses are stored. Of course this rights information should also be stored for the long term.

In the dictionary, all elements are defined and a rationale is given, together with a description and clarifying examples. Despite these explanations, a special group is now focusing on the implementation of Premis in repositories, as the transition of the data dictionary into a practical data model of the institution is not an easy task.[36]

6.5.5 Cost and creation of metadata

Although metadata are vital for long-term preservation of digital material, the creation of metadata is a costly business. Adding qualified metadata requires special skilled professionals and checking of these metadata; for example the creation of bibliographic metadata cannot always take place automatically. The perfect moment to collect metadata is when the objects are created. Several software programs automatically add metadata like size, creation, software and version. They also give the author the possibility to add metadata manually via the 'Document Properties' where the author name, title, number of pages and so on can be described. Despite the possibilities to create these metadata, not all authors are disciplined enough to add these metadata to the document. Guidelines by the repository managers can support the author and stimulate the creation of metadata. There are several metadata schemes; some are better suited to support certain digital material than others. The repository must have personnel that is qualified to compare the different metadata schemes in order to make the right decision, and to offer support to the author. Unfortunately, there is no optimal metadata scheme for all digital objects yet.[37]

6.5.6 Extracting metadata

Enormous efforts have taken place in automating the extraction of metadata as much as possible. Some programs make it possible to extract (mainly technical) metadata. One example is JHOVE (see 6.3.6), that can extract

different characteristics, depending on the file format. It is up to the repository manager to decide what to do with this extracted information and to which level of detail the information needs to be extracted. One option is to extract all the technical metadata and to store this information with the object, leaving the evaluation of the information to next generations. Another option is to evaluate the information before storing the object. Error codes will then be studied and a selection of the metadata will be made for permanent storage. The JHOVE output can be quite large if the repository does not make a selection. The national library of New Zealand developed a tool that extracts a set of preservation metadata from the header of a range of file formats, the 'metadata extraction tool'.[38] The output is captured into an XML file, which can be archived with the object.

6.5.7 Storage of metadata

Collecting metadata is not a static process. Although a set of metadata can be captured when the object is created or archived into the system, due to its long-term preservation it is likely that over the years some extra metadata will be added in order to preserve the digital object. Events like a migration of the object or the recalculation of a digital signature may lead to adding new metadata. As standards in this area are evolving, it is expected that extra metadata will be added during the lifetime of a digital object. This requires a flexible design of the repository that can facilitate these actions properly, and procedures to add these metadata. Metadata that are stored together with the AIP in the archive, make the update of the metadata more complex, in contrast with metadata stored in a separate data management part of the archive.

Apart from the storage of metadata, the institutional repository should make a selection of which metadata to store and where. Representation information, or metadata on how to render an object, might be stored together with the object. However, in several cases registries of metadata (for example on file formats and their representation information) would be preferable above individual initiatives of institutions. The aforementioned PRONOM and GDFR registries are new initiatives in this area.

6.6 Preservation and permanent access strategies

6.6.1 Background

In the ideal situation software and hardware will always work perfectly over the years, together with the created data. Yet this is not the real situation. Everyone knows that software releases are quickly launched and that hardware suppliers are eager to bring new and better devices every month. It is not always of commercial interest for the supplier to ensure that the data

are still usable with the new software. If they do support this, the result can be that certain original features are no longer supported, or that functionalities disappear. Sometimes software is phased out and has no successor; consequently original data files are not usable any more.

If we want to preserve data for the long term, we need to be proactive to avoid that data become inaccessible. These actions are called 'permanent access strategies'. In the previous chapters, advice has been given on how to collect relevant information for preservation to be used as accompanying metadata with the object. In this chapter, the strategies about how to deal with the software and hardware that run the programme, using the digital objects, will be the topic. Basically there are the following preservation strategies:

- technology preservation: preserve the original environment in the computer museum;
- migration: change the digital object and make it fit for a new environment;
- emulation: change the environment and make it fit for the unchanged object.

Up until now there is not a single preservation strategy that holds the ultimate solution for all digital objects. Institutional repository managers therefore should make a choice in order to be able to prepare the objects in the best way (metadata, chosen file format etc). This decision is highly dependent on three factors: the designated community (or the future user), the costs and the characteristics of the material.

Designated community

The designated community will have expectations towards the digital material and might wish to see the objects in present-day format or in the original look and feel. Both choices require another strategy of preservation. In the first case, migration might be an option while in the second case emulation will be preferred.

Characteristics

For the preservation strategies, it is important to realise which significant properties a digital object possesses and which of those properties should be saved for posterity. Five elements are seen as indicators for characteristics: readability, comprehensibility, appearance, functionality and look and feel.

If you can define the characteristics, you can capture them in metadata and the future generations can honour them while rendering. This exercise of defining the characteristics is a task for the repository itself.[39]

In order to support the process of choosing the relevant preservation strategy, a group of researchers at the Vienna University of Technology designed a conceptual framework for preservation planning. A process of several steps helps setting the criteria. Combined with a set of tools this will result in an outcome of the relevant preservation strategy and advice on the

tools to be used. This framework will be released as part of the European project Planets (2006-2010).

6.6.2 The computer museum

In several places worldwide people are collecting hardware and keep it alive and working. Although this might help in urgent cases of saving material under threat, this strategy of technical preservation is often seen as too fragile to use as a preservation strategy. Computers are made of material that deteriorates and often the basic elements are not replaceable, resulting in a dysfunctional machine.

6.6.3 Migration

Migration is the preservation strategy whereby the digital object is changed to make it accessible in a new environment. Although migration is a widely accepted preservation strategy, there are some drawbacks. Firstly, once started with migration, the repository will need to perform this action again and again over the years. But an error, once introduced with a migration, might enlarge with new migrations, perhaps leading to a damaged or inaccessible object. Another point is that testing of the results is difficult, time consuming and not yet automated. Migration requires thinking about the characteristics of the objects and it is highly unlikely that after migration all characteristics are still available in the new object. In international publications on the subject, the following migration methods are mentioned:

- Migration from one particular file format to a newer version of the same format. This process needs to be repeated whenever a new version is brought out. Research on this type of migration however, showed that more errors appeared when migrating from one version to the next version of software, then if the migration procedure skipped some versions.[40]
- Converting a file from one particular file format to another type of format. For instance, several institutions are known to convert files to formats more suitable for archiving purposes. This process is called 'normalisation' and is mainly carried out before a file is archived. For normalisation, a new file format PDF/A is thought suitable. However with the growing possibilities of file formats PDF/A might not be able to do right to the original documents as there is a restriction of possibilities.
- Batch migration, or converting a file from a particular file format into another type of format after the digital object has been archived.
- Migration on request: no periodical migration is carried out in this case, but individual files are migrated on demand. LOCKSS and the CAMiLEON project both did a test with this type of migration and concluded that it was a usable strategy for long term preservation.[41]

6.6.4 Emulation

Emulation is the process of bringing digital objects back to life in their original environment on top of a different computer environment. This process is carried out by an emulator, which is, as phrased by the Digital Preservation Testbed 'a program that runs on one computer and thereby virtually recreates a different computer'.[42] In this definition the word 'virtual' denotes that the emulator functions like the original computer, but differs physically. The original computer is called the 'target platform', the computer that executes the emulator is called the 'host platform'.

Emulation can be done at three different levels: the application software level, the system software (operating system) level and the hardware level. Emulating both application and system software requires knowledge of their design and implementation. These products are complex and very often proprietary, which makes it difficult to emulate. Another issue with application level emulation is that each application requires a specific emulator.

Emulation can also be done by mimicking the hardware architecture through software, which is called *software emulation of hardware* or *full emulation*. In this way, computer hardware, like a processor, is emulated by a software surrogate. In fact, the emulator itself is a software application which is placed on top of the operating system (OS), instead of running directly on hardware (see figure 14).

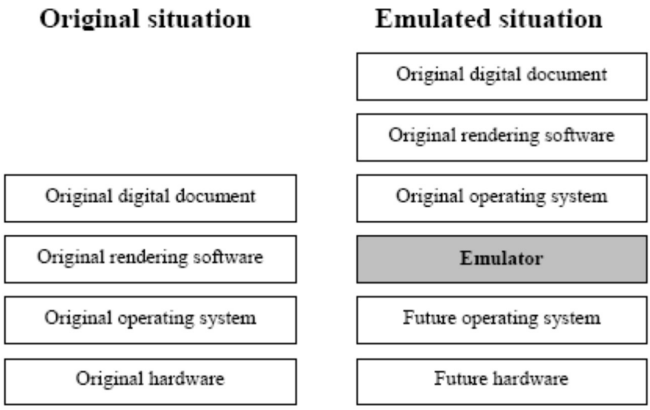

Figure 14 – Emulation of computer hardware using an emulator[43]

Although full emulation can be quite complicated, it has a more straightforward behaviour than emulating higher levels. For example, higher-level features, such as graphical user interfaces and running multiple applications side-by-side (known as multithreading), are difficult to emulate accurately. Emulation of a hardware platform does not incorporate these aspects, but requires the reproduction of the functional behaviour of the original platform in such a way that the original software is not able to distinguish the

difference between emulation and reality. Because hardware specifications are well defined and most often available, this behaviour is easier to reproduce than that of an OS or software application. Moreover, this approach retains the original OS, applications, drivers and configuration, which secures better authenticity of the original environment. This is the reason why most emulators are focused on recreation of hardware functionality.

Pros and cons

Since emulation is proposed as preservation action in the field of digital preservation, it has been subject of debate. Advocates of it propose it as the only solution for a large class of digital objects, although opponents define it as too complex and costly. Not having the final word said, with today's knowledge some benefits and disadvantages may seem clear (this was also addressed in an official statement on emulation during the Emulation Expert Meeting 2006 held in The Hague 20 October 2006)[44]:

Advantages

- It may be the only viable approach to preserving digital artefacts that have significant executable and/or interactive behaviour, like websites and multimedia applications.
- It does not require specific knowledge of every existing file format structure and properties.
- It presents the user the original look and feel of the digital object.
- One single emulator (plus the necessary software) can preserve digital objects in a vast range of arbitrary formats without the need of regular migrations.

Disadvantages

- Building an emulator is quite complicated which introduces high initial costs.
- Apart from the emulator, the original software environment needs to be preserved as well. This requires a dedicated (preferably worldwide) software repository, together with all other dependencies (fonts, DLL files, etc).
- Certain skills of installing and using the original software environment are required. Therefore, it is important to take notice of old manuals, installation instructions and computer workflow.
- As each emulator itself relies on an underlying computer platform, the interface between emulator and host platform should be kept operational over the years. Although certain solutions are proposed (i.e. using a virtual machine) it always requires a permanent amount of maintenance.

Current status

A lot of commercial and non-commercial emulators have been developed to achieve better use of hardware facilities. Virtualisation shows the benefits of more flexible computer environments. Furthermore, emulation/virtualisation is used for creating backwards compatibility with software that relies

on older types of hardware. So although emulation is not new in the world of information technology, it is in digital preservation.

With the scope of digital preservation, a few initiatives were started based on emulation. From 1999 until 2003 the CAMiLEON project conducted research into the possibilities of emulation as a digital preservation strategy.[45] Furthermore, in 2005 the KB, the National Library of the Netherlands[46] and the National Archive of the Netherlands started a joined emulation project. This project focuses on developing the first x86 emulator for digital preservation with strong requirements on flexibility and durability. This open source emulator was delivered in 2007.[47]

6.6.5 Universal Virtual Computer [48]

Together with IBM Netherlands, the Dutch National Library has developed a new preservation strategy, based on the Universal Virtual Computer (UVC). With the UVC it is possible to read files without adapting them and without the original hardware or software.

How it works

Every computer file can be revived with the UVC-based preservation method. Text documents, sound samples, images, spreadsheets or videos can all be reconstructed if a UVC is available for that particular type of format. The concept of the UVC was developed by IBM researcher Raymond Lorie.

Analogous to today's computer architecture, the UVC is a virtual representation of a simplified computer. Due to its simplicity, the UVC can in fact be made to work on any conceivable computer system. Basically, an extra layer on top of ever-changing hardware and software is created, which offers a stable platform to UVC programmes. Detailed instructions will enable future developers to rebuild a UVC at any given time.

Four components form the basis of the UVC

The **format decoder** is a programme developed for the UVC, by which a particular file format can be deciphered into a so-called **Logical Data View** (**LDV**). This LDV describes in detail how the digital object is structured. A **Logical Data Schema** (**LDS**) determines which elements can occur in a particular format and how these are related. For instance, raster-based images are defined by pixels and each pixel is composed of red, green and blue. Furthermore, the LDS contains information on the semantics of the different elements. What exactly is the colour blue and how can this meaning be captured so future users can see the authentic colours? This type of information is described in an LDS, one of which has to be made for every file format. Finally, the LDV is translated into an understandable representation by a **viewer**, the fourth and final component of the UVC-based preservation method.

Evaluation of the method shows the UVC to be a promising technique. JPEG and GIF images can be reconstructed in the future with the UVC-

based preservation method. However, the method needs to be elaborated. Decoders, LDS and viewers must be developed to make the UVC suitable for a wider range of digital material.

6.6.6 Conclusion

Preservation starts with creation, but the actual use of a preservation strategy requires also some extra work before storing the material, like information on characteristics, metadata and file format information.

6.7 Organisational aspects of digital preservation

As digital preservation is a commitment for an indefinite length of time, the organisation starting with this activity should realise what tasks and responsibilities are involved in this task. Institutional repositories are often started with enthusiasm and a limited budget, but it is wise to think who will take care of the collection in the repository if the institution can no longer afford it.

Apart from the OAIS model, which is a conceptual framework, there are existing standards regarding information technology and records management. But as yet there are no standards describing the requirements of a trusted digital repository and the organisation that runs it.

6.7.1 Checklists and audits

Several initiatives started defining these requirements of trusted digital repositories. In 2002, the former RLG (Research Libraries Group) and OCLC jointly published *Trusted Digital Repositories: Attributes and Responsibilities*, which articulated a framework of attributes and responsibilities for trusted, reliable, sustainable digital repositories capable of handling the range of materials held by large and small cultural heritage and research institutions. The next step was the idea of auditing these repositories. In 2003, RLG and the US National Archives and Records Administration created a joint task force to specifically address digital repository certification. The goal of this task force has been to develop criteria to identify digital repositories capable of reliably storing, migrating, and providing access to digital collections. The challenge has been to produce certification criteria and delineate a process for certification applicable to a range of digital repositories and archives, from academic institutional preservation repositories to large data archives and from national libraries to third-party digital archiving services. This initiative resulted in 2005 in a draft version of *An Audit Checklist for the Certification of Trusted Digital Repositories*, the final version of which was published in 2007.[49]

The German Nestor group produced their draft for public comment of the *Catalogue of Criteria for Trusted Digital Repositories*[50] in 2006. The Digital Curation Centre in the UK is also working on audit criteria. All three documents have taken the OAIS as a starting point, using the concepts and terminology of OAIS to define the measurable criteria. These initiatives are working closely together and one can expect that the different audit checklists eventually will result in an ISO standard. But regional and national differences (different laws, different financing models for public institutions etc.) will require national variations.

DRAMBORA[51] (Digital Repository Audit Method Based on Risk Assessment) was announced in 2007. Developed by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE), the toolkit guides organisations to perform internal audits and so to be well prepared for an external audit.

The checklists will not only be used by officials performing the audit (it is still undefined who will be capable of performing these audits), but are also meant as a guide for institutions to set up a repository, and to help the existing repositories to do a self-assessment of their repository. Even for repositories who outsource their activities, the audit lists can be of help to define their requirements and to judge how they are met.

It is important for a repository that its activities and intentions are transparent and well documented, so that the people who deposit their material know what to expect. This is also valid for the designated community of the repository. To give an impression of the criteria the audit checklists use, I will take the RLG/NARA[52] audit checklist as an example to show on which areas the audit focuses. The RLG/NARA audit checklist focuses on four areas:

- organisation
- repository functions
- the designated community and the usability of information
- technology and technical infrastructure

6.7.2 Organisation

Several aspects of the organisation around the repository are considered very important. It starts with the mission statement of the repository, whether this reflects its governance and viability, and if there is enough well-qualified staff to support this mission. Have measures been taken like escrow and a contingency planning in case something goes wrong? Is there a programme for constantly training the staff as developments in digital preservation evolve? And are the procedures and workflows that underpin the mission of the repository organised and well documented? It is important that the repository is financially well sustained. Even with effective business planning, somewhere in the future it might happen that resources are exhausted: did the repository foresee this and did it take sufficient measures in case this happens? There should be a succession plan, in

order to guarantee the existence of the repository in case some extreme situation happens. And are the copyrights, licences and liabilities well defined and documented? Especially in the area of institutional repositories, with e-prints, this is a tricky area. Authors might have signed contracts with publishers which do not allow them to store the article in the repository or elsewhere, not even as a preprint. It is the responsibility of the repository to deal with this topic and to be aware of the risks of the content of its repository. There is a strong movement nowadays to have all government-financed scientific output be open access, but this discussion does not cover all material and has not yet a final conclusion. Apart from access to the object, copyright issues might prevent the repository to take digital preservation strategies like migration and normalisation, which is a serious threat to the long-term preservation of the digital objects.

6.7.3 Repository functions

In this area the OAIS functions are followed and special requirements are formulated around Ingest, Archival Storage, Preservation Planning, Data Management and Access Management. The ingest procedure requires that the ingested material is in accordance with the mission of the repository, so that the (future) users should understand the relation between the repository and the material ingested. The depositor should have a clear idea about how the delivered package will be changed into a Submission Information package (for example, which checks are done and which metadata are added to the package of the depositor), and later into an Archive Information Package, (AIP), so that the depositor will know what will be preserved. It is important for the repository to have the right to perform long-term preservation actions, for example the right to migrate the AIP to another format, so this should be part of the contract with the depositor.

A critical component of any repository is its data management functionality. Regardless of technical composition the system needs to be able to store and use descriptive information (metadata) for access and retrieval. Descriptive information in this sense includes more than the narrative description that might be familiar to the user of a traditional library or archive catalogue. It also includes technical information necessary to preserve and manage the object. For example, if the collection of the repository contains objects with folk music of native people, this might require a sound card and high-quality speakers (in contrast for example to a digital object on which there is spoken text only). Preservation metadata will describe more in detail the requirements for reliable rendering of the object. These requirements from audit perspective ask for well-skilled and trained personnel and a management that is responsible for the fulfilling of its mission

6.7.4 Designated community

For the designated community it is important to know what they can expect from the repository, what they can ask for, when and how, and what the costs are. The repository should clearly state what of the digital object will be preserved. In some cases the repository might make a distinction in several classes of objects with different preservation levels, but this should be clear to the designated community. The data management functionality, where the metadata is stored, is critical for the designated community, as this is the place where the information for access and retrieval is stored. This is more than just bibliographic information like in the library catalogue. The character of the digital document requires that information is also stored about the technical requirements to preserve, manage and render the object. The repository owner should guarantee that enough metadata in this area is stored and, if necessary, over the years extended with new crucial information. 'In order to adequately provide digital preservation services, the repository must state its assumptions about the intended use of the *information objects* (i.e. content information and PDI) it will hold and preserve. The assumptions provide the foundation of the scope of services required to satisfy the information needs of the users of the collections in the repository. Without this foundation the repository cannot state the boundaries of its expected capabilities'[53].

6.7.5 Technology and technical infrastructure

This area deals with the technical infrastructure and the special technical requirements needed for long-term storage of digital objects. It does not prescribe which software to choose, it only gives the outline and requirements of the technical infrastructure. This section is based on the ISO standard 17799 for computer practices and is divided into three areas: system infrastructure requirements, the use of appropriate technologies for its designated community and security. As stated in the prefix of section 6.7 it is likely that repositories that perform 'good computing practices' will meet the requirements.

The various audit lists offer good guidance for every owner of an institutional repository but it is good to keep in mind that the checklists are working documents that will be constantly revised and augmented, as the practice of digital preservation is still evolving.

6.7.6 Research on costs of digital preservation

As everyone realises that digital preservation is a task that costs money, it is difficult to predict how much it will cost for an institution because several aspects play a role, for instance, what kind of digital objects are stored,

which preservation strategy will be chosen (migration, emulation), which tools are available to perform preservation actions, etc.

The LIFE project[54] from the British Library developed a formula about the costs of digital preservation. In this model the life cycle costs for a digital object over a certain time consist of a summary of the costs for Ingest, Metadata, Access, Storage and Preservation. Based on three case studies, the project could make a rough estimation of the costs of preservation. In this model, the costs for preservation (apart from the costs for acquisition, metadata, access, storage and ingest) consist of technology watch, preservation tool costs, preservation metadata, preservation action and quality assurance. The overall conclusion of this project is that costs are measurable. The report also concluded that collaboration in preservation tool development is necessary to bring the costs down for all preservation institutions. This project will be continued.

The ESPIDA project[55] is developing 'a sustainable business-focused model for digital preservation at a Further/Higher Education institution'. It will bring digital preservation to the core of strategic thinking, planning and culture at the University of Glasgow and will disseminate the model to the wider community.

Outsourcing of digital preservation for institutional repositories might be an option for some institutions, as a lot of them will not be able to perform the task of digital preservation adequately due to lack of staff, budget and time. A serious investigation of the advantages and disadvantages of this solution was done in the SHERPA project that was organised to set up a network of shared responsibility for institutional repositories, SHERPA DP.[56]

6.7.7 Cooperation and preservation watch

Digital preservation is not an activity one repository can fulfil on its own. It requires permanent research, as the digital world is constantly changing. It is important to keep up with the developments everywhere in the world. Several initiatives exist where an overview of these activities is regularly updated[57] and different digital journals about all aspects of digital preservation are regularly published.[58] International conferences offer an up-to-date view of the latest developments. As long as digital preservation is not a matter of 'just reading the manual', it is important to keep oneself informed, in order to make the right decisions for this precious material. So join the digital preservation community!

Appendices

1 Roadmap of initiatives on Intellectual Property Rights

Austria

The Austrian Academy of Sciences Press operates the institutional repository of the Austrian Academy of Sciences. The press is both a publishing house and an institutional repository. Publications by scientists of the Austrian Academy are uploaded as far as external publishers involved give the institutional repository permission to do so. The Austrian Academy only uploads publications to the repository on the basis of a general agreement that has to be signed by the director of the institute before the upload to the repository starts. The Austrian Academy of Sciences Press developed a user guide (available in German) at http://www.epub.oeaw.ac.at/dokumentation. The contact person is Herwig Stoeger, herwig.stoeger@oeaw.ac.at.

The central body in Austria for the funding of basic research is the Austrian Science Fund Organisation (FWF). The FWF is a signatory of the Berlin Declaration and committed to the support and promotion of open access to scientific publications. Therefore the FWF expects the results of research it supports to be made publicly available free of charge. Researchers participating in FWF projects should attempt, as far as possible, to secure lasting and non-exclusive rights for the electronic publication of their results, for the purpose of non-profit-oriented utilisation in their contracts with publishing houses. If there is an embargo period, this period should last no longer than six to twelve months. The FWF reimburses the publication costs.

If the results of projects are published in a conventional peer-reviewed journal, an application for reimbursement of costs associated with the submission of scientific articles to refereed Open access journals may be sent to the FWF. The FWF has a website with information about open access. There is a small paragraph regarding copyright aspects; it refers to the Creative Commons website and the RoMEO e-prints site. The website also points to the Open access publishing models of Blackwell, Oxford University Press and Springer.

The URL of the web site is http://www.fwf.ac.at/en/public_relations/oai/index/html.

Contact person is dr. Falk Reckling, falk.reckling@fwf.ac.at.

Belgium

The department of research of the Université Libre de Bruxelles (ULB) is in charge of the legal framework for the relationship author-institution regarding intellectual property rights. The ULB defines the legal framework for

these rights in research results in the 'rules regarding property, protection and development of results of research performed at the ULB' ('Règlement en matière de propriété, de protection et de valorisation des résultats des recherches effectuées à l'Université Libre de Bruxelles') which can be found at http://www.ulb.ac.be/ulb/greffe/docs/reglement/regl-propri-recherche.html. The ULB identifies the author as the owner of the copyright of his publications. The website of the university's electronic theses and dissertation repository dealing with the legal aspects gives this information. It can be found at http://www.bib.ulb.ac.be/fr/bibliotheque-electronique/theses-bictele/aspects-juridiques/index.html.

The ULB has no institutional policy or recommendations to researchers regarding transfer of copyright, or recommendations to use specific copyright agreements. The library is setting up an institutional repository, but the legal aspects related to copyright on publications have not yet been dealt with at the institutional level. So far, the library checks the Romeo database for the publishers' policies regarding self-archiving. The library of the ULB provides some information on copyright issues and recommendations about copyright transfer, as well as useful links to copyright contract addenda on its website http://www.bib.ulb.ac.be/fr/crise-de-la-publication-scientifique/les-droits-dauteur/index.html and http://www.bib.ulb.ac.be/fr/crise-de-la-publication-scientifique/et-vous-que-pouvez-vous-fainstitutional repository e/index.html. Contact persons are Marylene Poelaert, marylene.poelaert@ulb.ac.be and Francoise Vandooren, francoise.vandooren@ulb.ac.be.

The University of Namur has a research centre specialised in intellectual property, the Centre de recherche informatique et droit (CRID). This centre organises training sessions on Open access as a means to add value to research results. The website can be found at

http://www.fundp.ac.be/facultes/droit/recherche/centres/crid.

Denmark

In Denmark, copyright is vested in the author and it may be transferred to the employer either by contract or according to tradition. However, universities, public research institutions, public hospitals or public medical research institutions only have copyright for the employee's research output if there is an explicit agreement.

At the Technical University of Denmark (DTU), work on Open access was done on a local as well as a national level for several years. All necessary knowledge regarding copyright and repositories is collected on an intranet of the university. Researchers need to send a contract, designed by the DTU and based on a national contract, to the publishers. In this contract the exclusive rights are kept exclusively by the researchers and are not transferred to the publishers.

So far, according to the experiences of the DTU, few researchers have been willing to put up with the effort in keeping the copyright of the institutional repository publications. Deposition of journal articles is still not widely implemented. Theses are the main focus at the moment; journal

articles will follow later. DTU has three access levels from which the author can choose. The publisher's version is not deposited because the system only enables preprints, in- and post-prints. To find out whether uploading of the print is permitted, the SHERPA/RoMEO database is used. Contact person for Open access at DTU is Liv Fugl, lf@dtv.dk.

Estonia

By regulation 17 of 18 November 2003, the council of the University of Tartu approved the 'Principles governing intellectual property' at the University of Tartu. The university states that the economic rights in the result of creative activity of an author shall belong to the university. If the economic rights don't transfer pursuant to law the economic rights shall transfer from author to the university under an agreement or any other written document.

The economic rights of an author in works that have been created in the execution of the dinstitutional repository ect duties of the author, belong to the university, in accordance with Estonian copyright law. With regard to scientific articles and conference proceedings, the university waives the economic rights in favour of the author. This is recorded in an employment contract or any other agreement between author and university.

France

Every week Archimer, the institutional repository of Ifremer, checks the publications that submitted for storage in the repository. The publications edited or coedited by Ifremer are checked in the database Current Contents Connect®. Then for each publication the self-archiving policies of publishers are checked on the SHERPA/RoMEO website. If the publisher's policy does not allow storage, Archimer systematically tries to contact the editor and get permission to upload the publication in the repository. Contact person is Frederic Merceur, Frederic.Merceur@ifremer.fr.

Germany

The Stuttgart University Library and the Computer- und Medienservice of the Humboldt University in Berlin jointly carry out a project, sponsored by the Deutsche Forschungs Gemeinschaft (DFG), to answer the question of what is permitted by German publishers regarding publication of a scientific paper through an institutional repository. Theinstitutional repository aim is to integrate information about German publishers into the SHERPA/RoMEO list, and to provide a German-language interface for this list. Furthermore the project brings together the experiences of German libraries DINI (www.dini.de), and it provides an infrastructure for dinstitutional repository ect inclusion of the SHERPA/RoMEO list into the institutional repositories. More about the project can be found on http://www.ub.uni-stuttgart.de/oa-policies and Klaus Wendel can be contacted for more information, klaus.wendel@ub.uni-stuttgart.de.

The co-ordination office of the Helmholtz Association is working on the implementation of Open access in its fifteen research centres. With an an-

nual budget of 2.3 billion euros, the Helmholtz Association is the largest research institution in Germany. Since its centres are (still) very independent, theinstitutional repository efforts for Open access very widely. Most of the centres have up and running institutional repositories, but no explicit copyright policies. One of the exceptions is the Forschungszentrum Jülich. Juelicher Wissenschaftliche Elektronische Literatur (JUWEL) is the official research centre's institutional repository. In JUWEL, publications of the centre's scientists are filed, preserved, made accessible and distributed. The JUWEL website has a lot of information about open access, including copyright. The centre advises its authors to add on the following clause in theinstitutional repository agreements with publishers: 'Forschungszentrum Jülich GmbH shall be entitled to make the article freely accessible to the general public through the institutional repository of Forschungszentrum Jülich at the time of publication (alternatively three or six months after publication of the article)'. The information can be found at http://www.fz-juelich.de/zb/oa_modell#. Contact person is Cornelia Plott, c.plott@fz.juelich.de.

The research project 'Open access Recht' of the legal faculty of the Georg-August University of Goettingen supports authors who want to publish in Open access journals. The book 'Rechtliche Rahmenbedingungen von Open access'[1] gives an introduction in the Open access publication model and answers the legal questions concerned with this model.

Netherlands

IGITUR, the digital repository of the University of Utrecht, has developed a deposit licence. It is built into theinstitutional repository repository. By clicking on the option 'Grant the licence', the author gives IGITUR permission 'to store the document in the academic digital archive and make it available on the internet'. The IGITUR website is http://www.igitur.nl/en/default.htm. IGITUR's publisher advisor is Astrid van Wesenbeeck, info@igitur.uu.nl. In 2006, Leiden University's library opened its copyright information website. This website assists authors to make theinstitutional repository publications as visible as possible. There is a FAQ as well as information about copyright policies of publishers.

Norway

As yet, only the University of Oslo has decided to make deposition in a digital repository compulsory. From 2007 it will be mandatory for all postgraduate students at this university to submit theinstitutional repository theses electronically. Authors in Norway are assisted by the NORA project (Norwegian Open access Archives). NORA is establishing an Open access window including information for authors on what to do in order to retain the right to self-archive an article. The NORA Open access window is complementary to ScieCom in Sweden and the SHERPA list.

Sweden

Only two universities in Sweden have a policy regarding open access: Lund University and Stockholm University. In Sweden a programme called ScieCom, Swedish Resource Centre for Scientific Communication, was run by the National Library of Sweden. The aim of the project was to provide information about present developments in scientific communication, and to promote Open access to scientific publications. Information was communicated via the website of the project, through seminars and presentations, and via an Open access e-journal, ScieCom Info. The project was led by Lund University libraries. More information can be found at http://www.sciecom.org/.

The National Library of Sweden is currently supporting a project called 'Copyright in a new publishing envinstitutional repository onment'. The project strives to produce up-to-date, practical and easily understandable information about copyright in connection with scientific publishing. It will investigate current practice and interpretations at Swedish higher education institutions, present instructive international examples, and discuss relations between author and institution, author and publisher and law and contract. Information will be presented via a website, as well as via seminars and courses. The project is led by Ingegerd Rabow at Lund University Library, ingegerd.rabow@lub.lu.se.

United Kingdom

As part of the Clearing the Way research project supported by the 2000 Elsevier/LINSTITUTIONAL REPOSITORY G Research Award, an A-Z list of the UK's higher education institutions' copyright pages was composed. This list can be accessed through http://www.lboro.ac.uk/library/skills/crightpages.html. Most of the pages are intended to provide guidance on matters of copyright and the copying of material for research, teaching and learning at universities. Sometimes the copyright policy of a university can be read, hidden on one the pages.

Binstitutional repository kbeck University London developed the Binstitutional repository kbeck ePrints Deposit Licence.[2] This licence is designed to give repository administrators the right to store, copy and manipulate the material in order to ensure that the material can be preserved and made available in the future. It also confinstitutional repository ms that the depositor has the rights to submit material to the repository. The licence is non-exclusive and the author retains his rights. The Binstitutional repository kbeck repository has the right to distribute electronic copies of the work for the lifetime of the repository or an agreed time span. The repository has the right to translate the material to ensure it can be made accessible in the future. The author has the right to remove the work at any point in the future, just as the repository has. The metadata record that the work was stored will remain visible for the lifetime of the repository.

Bristol University has the ROSE deposit licence.[3] The depositor grants a non-exclusive, royalty-free licence to the university's information services

department on behalf of the university. The information services department then has the right to make copies of the work available for distribution worldwide in an electronic format and in any other medium or format for the lifetime of the project or for the purpose of free access without charge. The repository is also allowed to electronically store, translate, copy or rearrange the work to ensure future preservation and accessibility. The repository can incorporate metadata or documentation for the work in public access catalogues. The University of Nottingham supports its staff in complying with mandates of funding organisations and recommendations. Information can be found at http://eprints.nottingham.ac.uk/guidance.html#journalrules.

2 Additional reading on business models

There are a number of reports, articles and other resources available on developing and running a repository and this section presents a quick overview of some of them.

Repository surveys and overviews

Two studies have been published recently on the repository state of play in the United States. Although strictly outside the sphere of interest of DRIVER, which is a European project, nevertheless these studies do provide some helpful context from across the Atlantic. In July 2006 the Association of Research Libraries (ARL) published the results of a survey of 87 member libraries asking about their repositories or plans for repositories. The survey, one of the ARL's SPEC Kit series, is detailed and informative, collates the responses from these institutions in a usable way, and covers many issues from staffing, technology, costs, and policy developments through to marketing and advocacy:

University of Houston Institutional Repositories Task Force (2006) *SPEC Kit No. 292: Institutional Repositories.*http://www.arl.org/resources/pubs/spec/complete.shtml.

The Council on Library and Information Resources published a census of US institutional repositories in February 2007. This is the first phase of the MIRACLE project and is a census to determine the involvement of US institutions with repositories. A further four phases of the project will be published in time.

Karen Markey, Soo Young Rieh, Beth St. Jean, Jihyun Kim, and Elizabeth Yakel, *Census of Institutional Repositories in the United States: MIRACLE Project Research Findings* (web only publication, 2007) http://www.clir.org/pubs/abstract/pub140abst.html.

Two years ago, a similar exercise, though global in scope, was carried out for the CNI-JISC-SURF conference on *Making the Strategic Case for Institutional Repositories* and the results were published by SURF:

Gerard van Westrienen, *Country update on academic institutional repositories* (Utrecht, 2005) http://www.surffoundation.nl/download/country-update2005.pdf.

In mid-2006 a review of institutional repositories with a focus on information systems was published from Australia:

Mary Ann Kennan, and Concepción Wilson, "Institutional repositories: review and an information systems perspective", *Library Management* 27 (4/5) (2006) http://dlist.sir.arizona.edu/1200/

Handbooks and guides

The Open Society Institute funded a detailed and helpful guide to repositories and self-archiving from Southampton University's School of Electronics & Computer Science:

Leslie Carr,, *EPrints Handbook* [online text] (2003) http://www.eprints.org/documentation/handbook/.

The generously funded DSpace repository at MIT has been the focus of much interest. Two of the staff have produced a guide to creating a repository.

Mary Barton, and Margaret Walters, *Creating and Institutional Repository: LEADIRS Workbook* (2004) http://www.dspace.org/implement/leadirs.pdf

Repository frameworks and landscapes

Walters described the infrastructural background to repositories:

Tyler Walters, 'Strategies and frameworks for institutional repositories and the new support infrastructure for scholarly communications', *D-Lib Magazine* 12/10 (2006). http://www.dlib.org/dlib/october06/walters/10walters.html.

A JISC-funded study on a national model for delivering open access articles to the UK research community proposed a harvesting model based on distributed institutional archives:

Alma Swan, Paul Needham, Steve Probets, Adrienne Muir, Ann O'Brien, Charles Oppenheim, Rachel Hardy, and Fytton Rowland, "Delivery, management and access model for E-prints and Open Access journals within further and higher education", Report of a JISC study (2004) http://eprints.ecs.soton.ac.uk/11001/

Another JISC-funded study looked at the potential for linking UK repositories and creating services to support them or to work from them. The report contains a section on business models for such services:

Alma Swan and Chris Awre, *Linking UK repositories: Technical and organisational models to support user-oriented services across institutional and other digital repositories: Scoping study report* (2006)

http://www.jisc.ac.uk/uploaded_documents/Linking_UK_repositories_-report.pdf

Appendix: http://www.jisc.ac.uk/uploaded_documents/Linking_UK_repositories_appendix.pdf

A book by Richard Jones, Theo Andrew and John MacColl gives a comprehensive account of institutional repositories:

Richard Jones, Theo Andrew and John MacColl, *The Institutional Repository* (Chandos Publishing: Oxford, 2006)

Accounts about specific repositories

A number of articles have been published over the last few years giving the perspective from individual repositories, and this is a selection of them:

Stephen Pinfield, Mike Gardner, and John MacColl, "Setting up an institutional e-print archive" *Ariadne* 31 (2002) www.ariadne.ac.uk/issue31/eprint-archives/intro.html.

William Nixon, "The evolution of an institutional e-prints archive at the University of Glasgow" *Ariadne* 32 (2002) http://www.ariadne.ac.uk/issue32/eprint-archives/intro.html.

Jörgen Eriksson, "More content in the institutional repository", *ScieCom Info*, 1 (2005)

http://www.sciecom.org/sciecominfo/artiklar/eriksson_05_01.shtml.

Jessie Hey, "Targeting academic research with Southampton's institutional repository", *Ariadne* 40 (2004) http://www.ariadne.ac.uk/issue40/hey/.

Studies that include information on economic aspects of repositories

As well as the cost of repositories given in the report by Swan et al (2004) above, this very detailed and informative report by John Houghton and colleagues delves deeply into the economic basis of scholarly communication as a whole (focused on Australia, but widely applicable):

John Houghton, Colin Steele, and Peter Sheehan, *Research communication costs in Australia: Emerging opportunities and benefits*, a report to the Department of Education, Science and Training. (Melbourne, 2006) http://www.dest.gov.au/NR/rdonlyres/0ACB271F-EA7D-4FAF-B3F7-0381F441B175/13935/DEST_Research_Communications_Cost_Report_-Sept2006.pdf

Notes

A DRIVER's Guide to European Repositories: Five studies of important Digital Repository related issues and good Practices (ISBN 978 90 5356 411 0) - DARE Repository: http://dare.uva.nl/aup/nl/record/260224

The European Repository Landscape: Inventory study into present type and level of OAI compliant Digital Repository activities in the EU (ISBN 978 90 5356 410 3) - DARE Repository: http://dare.uva.nl/aup/nl/record/260225

Investigative study of standards for Digital Repositories and related services (ISBN 978 90 5356 412 7) - DARE Repository: http://dare.uva.nl/aup/nl/record/260226

Notes to About the DRIVER studies

1. http://www.driver-community.eu/.

Notes to Introduction

1. The Open Archives Initiative Protocol for Metadata Harvesting http://www.openarchives.org/OAI/openarchivesprotocol.html
2. See the section "About the DRIVER studies" for more information about the DRIVER project.

Notes to The business of digital repositories

1. http://www.driver-repository.eu/.
2. http://www.darenet.nl/en/page/language.view/dare.start.
3. www.sherpa.ac.uk.
4. Chesborough and Rosenbloom, "The role of the business model" (2002).
5. Clarke, "Open source sorftware and open content" (2004).
6. Roosendaal et al., "Developments in scientific communication" (2001).
7. Roosendaal and Geurts, "Forces and functions" (1998).
8. www.arxiv.org.
9. Timmers, "Business Models" (1998), Rappa, "Business Models" (2000).
10. Swan and Awre, *Linking UK repositories* (2006).
11. http://hal.archives-ouvertes.fr/.
12. TARDis project final report : http://eprints.soton.ac.uk/16122/.
13. For an example of an e-portfolio repository see http://portfolio.ecs.soton.ac.uk/.
14. www.gla.ac.uk/espida.
15. http://www.oclc.org/research/projects/pmwg/; http://preserv.eprints.org/.
16. Fox, "The 5S Framework" (1999).
17. Swan, "The culture of open access" (2006), Sale, "Comparison of content" (2006).

18. See Registry of Open Access Repository Material Archiving Policies: http://www.eprints.org/openaccess/policysignup/.
19. Swan and Brown, "Open Access Self-archiving" (2007).
20. http://roar.eprints.org/; http://www.opendoar.org/.
21. Swan and Brown, "Open Access Self-archiving" (2005), Carr, "Use of navigational tools" (2006).
22. Swan et al., "Delivery, management" (2004), Houghton et al., *Research communication* (2006).
23. Adapted from Swan, "The culture of open access"(2006), and Sale, "Comparison of content"(2006)
24. Adapted from Swan, "The culture of open access"(2006), and Sale, "Comparison of content"(2006)
25. Carr and Brody, "Size isn't everything" (2007).
26. Moffat, "'Marketing'with metadata" (2006).
27. Lyon, "Dealing with data"(2007).
28. www.darenet.nl.
29. www.sherpa.ac.uk.
30. http://www.sherpadp.org.uk/; http://www.sherpa.ac.uk/romeo.php; http://www.sherpa.ac.uk/juliet/index.php.
31. See DAREnet list of services at http://www.darenet.nl/en/page/language.view/diensten.diensten.
32. http://www.darenet.nl/en/page/language.view/diensten.arnex.page.
33. http://lvmj.medfak.lu.se/.
34. http://iesr.ac.uk/.

Notes to The population of repositories

1. Crow, "The case for Institutional Repositories" (2002)
2. The Berlin Declaration on Open Access: http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html, launched in October 2003.
3. Harnad, "Self-archiving", 2006, Sale "The acquisition of Open Access", Sale, "The Patchwork Mandate" (2007)
4. Harnad, "Publish or Perish" (2006), .
5. Crow, "The case for Institutional Repositories", 16-17.
6. Mackie, "Filling Institutional Repositories" (2004)
7. Rowlands and Nicholas, "The changing scholarly communication landscape" (2006), 36.
8. Davis and Connelly, "Institutional Repositories" (2007).
9. Gierveld, "Considering a Marketing and Communications Approach" (2006).
10. Callan, "The development and implementation" (2004).
11. Crow, "The case for Institutional Repositories" (2002), 9.
12. Waaijers, "From libraries to libratories" (2005).
13. Day, "Institutional repositories and research assessment" (2004).
14. Crow, "The case for Institutional Repositories" (2002), 21. Sparks, JISC's Disciplinary Differences (2005).
15. Statement from the EUA Working Group on Open Access, EUA, 26 January 2007: http://www.eua.be/fileadmin/user_upload/files/newsletter/EUA_WG_open_access.pdf

16. Vanessa Proudman would like to thank the following interviewees: Eloy Rodrigues and Ricardo Saraiva (University of Minho), Daniel Charnay, Franck Laloë, Muriel Foulonneau, Francis André and Laurent Capelli (HAL), Jens Vigen and Joanna Yeomans (CERN), Leo Waaijers, Martin Feijen and Annemiek van der Kuil (Cream of Science), Wendy White, Jessie Hay and Les Carr (University of Southampton), Titia van der Werf and Ursula Oberst (Connecting Africa).
17. The Directory of Open Access Repositories – *Open*DOAR: http://www.opendoar.org/ ROARMAP (Registry of Open Access Repository Material Archiving Policies): http://www.eprints.org/openaccess/policysignup/
18. https://repositorium.sdum.uminho.pt/?locale=en
19. http://eprints.soton.ac.uk/
20. TARDIS: http://tardis.eprints.org/
21. http://cdsweb.cern.ch/
22. arXiv.org e-print archive: http://arxiv.org/
23. http://hal.archives-ouvertes.fr/index.php?langue=en
24. http://www.creamofscience.org/
25. http://www.darenet.nl
26. http://www.connecting-africa.net/
27. Sale, "The acquisition of open access research articles" (2006).
28. DOI—The Digital Object Identifier System: http://www.doi.org/.
29. Economists Online: http://www.nereus4economics.info/economistsonline.html
30. Massachusetts Institute of Technology (MIT): http://dspace.mit.edu/; Queensland University of Technology (QUT): http://eprints.qut.edu.au/
31. Harnad, "The Golden and Green Roads" (2003).
32. For a list of these research councils and their policies, see http://www.rcuk.ac.uk/research/outputs/access/default.htm
33. Netherlands Organisation for Scientific Research (NWO): http://www.nwo.nl/nwohome.nsf/pages/SPPD_5R2QE7_Eng
34. CNRS: Centre national de la recherche scientifique: http://www.cnrs.fr/index.html; Centre pour la Communication Scientifique Directe (CCSD): http://ccsd.cnrs.fr/
35. http://www.ascleiden.nl/
36. NOD—Dutch Research Database, http://www.onderzoekinformatie.nl/en/oi/nod/
37. Unfortunately, the few current download statistics available are too new to allow for much analysis of growth or take-up at various levels.
38. DSpace: http://www.dspace.org/
39. For example, that of INRIA: http://hal.inria.fr/
40. Economists Online: http://www.nereus4economics.info/economistsline.html
41. SOAP—Simple Object Access Protocol.
42. Figures are for 2005.
43. SPIRES http://www.slac.stanford.edu/spires/;
    arXiv.org e-print archive: http://arxiv.org/;
    RePEc: http://www.repec.org;
    PubMed Central: http://www.pubmedcentral.nih.gov/ or UK PubMed Central http://ukpmc.ac.uk/;
    OAIster: http://www.oaister.org/
44. BASE: http://base.ub.uni-bielefeld.de/index_english.html;
    INTUTE: http://www.intute.ac.uk/

45. For an overview of current services offered by the cases studied, see the www.driver-repository.eu
46. Davis and Connelly, "Institutional Repositories" (2007).
47. SHERPA/RoMEO: http://www.sherpa.ac.uk/romeo.php
48. Idem.
49. COMA website: http://kubl03.uvt.nl:4090/?request=coma&domain=Coma&-frame=dare
50. Simple Object Access Protocol, see http://www.w3.org/TR/2000/NOTE-SOAP-20000508/

Notes to Intellectual property rights

1. Article 5 Berne Convention for the Protection of Literary and Artistic Works.
2. Section 10 Copyright, Designs and Patents Act 1988.
   Provision L 113-2 of **Le code de la propriété intellectuelle** (Intellectual Property Code).
3. Article 12 Copyright Act 1912.
4. Article 13 Copyright Act 1912.
5. §11 Allgemeines of the Urheberrechtsgesetz vom 9. September 1965 (BGBl. I S. 1273), zuletzt geändert durch das Gesetz vom 10. November 2006 (BGBl. I S. 2587).
6. Telediffusion is the distribution by any telecommunication process of sounds, images, documents, data and messages of any kind.
7. Section 77 CDPA 1988.
8. Lucas.,*Traité de la propriétaire* (2006).
9. European Commission, "Green paper on copyright", *COM* (88) 172 final
10. European Parliament, "Directive 96/9/EC" (1996), 20–28
11. Fixtures marketing Ltd v Svenska AB C-338/02; Fixtures Marketing Ltd v Organismos Prognostikon Agonon Podosfairou EG C-444/02; Fixtures Marketing Ltd v Oy Veikkaus Ab C-46/02, British Horseracing Board Ltd v William Hill Organization Ltd C-203/02.
12. http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf
13. http://www.soros.org/openaccess
14. http://www.earlham.edu/~peters/fos/bethesda.htm;
15. Harnad et al., "The acces / impact problem" (2004).
16. Suber, "Open Access Overview"[website] http://www.earlham.edu/~peters/fos/overview.htm.
17. Idem.
18. http://www.sherpa.ac.uk/romeoinfo.html, http://users.ecs.soton.ac.uk/harnad/Hypermail/Amsci/0662.html http://users.ecs.soton.ac.uk/harnad/Hypermail/Amsci/0662.html.
19. http://www.alpsp.org/ngen_public/default.asp?ID=202&groupid=192&groupname=About+ALPSP
20. http://copyrighttoolbox.surf.nl/copyrighttoolbox/.
21. http://www.earlham.edu/~peters/fos/bethesda.htm.
22. Hess et all., *Open access and Science publishing*, (2007), 10.
23. http://copyrighttoolbox.surf.nl/copyrighttoolbox/authors/licence/.
24. http://www.sherpa.ac.uk/romeo.php.
25. http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/4296.html.

26. http://www.eprints.org/openaccess/policysignup/.
27. http://ec.europa.eu/research/eurab/pdf/eurab_scipub_report_recomm_dec06_en.pdf.
28. Brussels 14.2.2007 COM (2007) 56 final communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on scientific information in the digital age: access, dissemination and preservation {SEC(2007)181}.
    OECD, *Science, Technology and Innovation for the 21st Century. Final Communique,* (29-30 January 2004)-http://www.oecd.org/document/15/0,3343,en_2649_201185_25998799_1_1_1_1,00.html.
29. http://www.nih.gov/about/publicaccess and http://www.rcuk.ac.uk/access/statement.pdf.
30. http://www.rcuk.ac.uk/aboutrcuk/default.htm.
31. http://www.ivir.nl/legislation/nl/copyrightact.html.
32. http://www.iuscomp.org/gla/statutes/UrhG.htm#43.
33. Provision L111-1 of **Le code de la propriété intellectuelle** (Intellectual Property Code).
34. http://www.opsi.gov.uk/acts/acts1988/Ukpga_19880048_en_2.htm#mdiv11.
35. Mossink, *Auteursrechten* (1999).
36. Mossink, *Report on Institutional Copyright Policies* (2006).
37. Crews, *Copyright, Publishing and Scholarship*. (2007).
38. http://www.surf.nl/copyright.
39. http://www.surf.nl/copyrighttoolbox.
40. http://www.lboro.ac.uk/departments/dis/disresearch/poc/pages/pub-listing-rights.html.
41. http://www.dipp.nrw.de/lizenzen/dppl/index_html/dppl/DPPL_v2_en_06-2004.pdf.
42. http://www.ifross.de.
43. Hirtle,"Author Addenda" (2006).
44. http://www.sparceurope.org/.
45. http://www.arl.org/sparc/author/docs/AuthorsAddendum2_1.pdf.
46. Gareth Knight 21 June 2004 http://ahds.ac.uk/about/projects/sherpa/report.htm.
47. http://www.dare.leidenuniv.nl/index.php3?m=23&c=163&garb=0.5108627787361215&session.
48. Barker, The Common Information Environment (2006), 77.
49. Korn, "Creative Commons Licences" (2006).

Notes to Data curation

1. http://en.wikipedia.org/wiki/Curator, 12 February 2007
2. See also chapter 6 on Long-term preservation
3. The websites of the mentioned projects and initiatives can be found at: http://planets-project.eu, http://www.casparpreserves.eu, http://www.dpc.delos.info, http://www.digitalpreservationeurope.org, http://www.dariah.eu, http://www.clarin.eu, http://http://www.erohs.org, accessed February 2007.
4. "Digital Curation: digital archives, libraries and e-science seminar" sponsored by the Digital Preservation Coalition and the British National Space Centre held in London, October 19th 2001, http://www.dpconline.org/graphics/events/digi-

talarchives.html, accessed February 2007
Beagrie, "Digital curation for science" (2006).
5. http://www.ijdc.net/ijdc/ , accessed February 2007.
6. Beagrie, "Digtial curation for science" (2006), 5.
7. Beagrie,"E-infrastructure strategy" (2007), 5.
8. Idem, 8.
9. The website of the Digital Curation Centre (DCC) contains state-of-the-art contributions on data curation issues and relevant skills, see: http://www.dcc.ac.uk/events/pv-2005/, accessed March 2007.
10. CCSDS, *Reference model* (2002). The OAIS reference model is more extensively covered in the chapter on long-term preservation.
11. A 'Designated Community' is defined by the OAIS reference model as 'an identified group of potential users of the archives' contents who should be able to understand a particular set of information'. A 'designated community' is multifaceted and decisions about what to preserve must take into account not only the needs of current users, but also those of users far into the future.
See Kaczmarek, et al., "Using the audit checklist" (2006); Dobratz and Schoger, "Digital repository certification" (2005).
12. Svenonius, *The intellectual foundation* (2000).
13. Tindemans, 2006, 84.
14. Heery, 2005, 13
15. Brogan, 2006, 17-18
16. http://www.darenet.nl, accessed February 2007.
17. The website of the DARELUX project can be found at: http://www.library.tudelft.nl/darelux/projectinformatie/index.htm (in Dutch). The "archeological data repository can be found at: http://edna.itor.org/nl/. The DARC project website can be found at: http://www.ascleiden.nl/Projects/Darc/. The e-Laborate prototype can be found at: http://www.e-laborate.nl/en/, accessed February 2007.
18. Hunter, "Scientific publication packages" (2006), 37.
19. Lagoze, "The ABC ontology" (2001).
20. The model can also be expressed as an RDF representation, making the model relevant for the semantic web.
21. Hunter, "Scientific publication packages" (2006), 44.
22. IFLA "Functional Requirements" (1998). This is an indication that the concepts expressed in the publication are increasingly part of the common knowledge on information organisation.
23. Hunter, "Scientific publication packages"(2006), 44.
24. Hunter, "Scientific publication packages"(2006).
25. Examples of these models mentioned by Hunter are the content standard for computational models – CCSM and the 'Scientific Publication Metamodel – SPM. See: Hunter, "Scientific publication packages (2006). .
26. Hunter, "Scientific publication packages" (2006), 37
27. Formed in 1995, the Council for the Central Laboratory of the Research Councils (CCLRC) owns and operates the Rutherford Appleton Laboratory in Oxfordshire, the Daresbury Laboratory in Cheshire and the Chilbolton Observatory in Hampshire. These world-class institutions support the research community by providing access to advanced facilities and an extensive scientific and technical expertise.
28. Sufi, "CCLRC scientific metatada model"(2004).
29. Sufi, "CCLRC scientific metadata model" (2004) , 16.

30. The data quality issues discussed in this section are based on the philosophy of DANS (Data Archiving and Networked Services). DANS is the Dutch national organisation responsible for storing and providing permanent access to research data from the humanities and social sciences. The website of DANS can be found at: http://www.dans.knaw.nl, accessed cited 15 february 2007.
31. Concerning open access to scientific data the 'Petition for guaranteed public access to publicly-funded research results' is of great importance. See: http://www.ec-petition.eu/, accessed February 2007.
32. The PRONOM online registry contains data about data file formats and their supporting software products. See: http://www.nationalarchives.gov.uk/pronom/, accessed February 2007.
33. The most important standard for descriptive metadata is the Dublin Core Element Set; http://www.dublincore.org, accessed February 2007
34. An important initiative to standardise the features of preservation metadata is the data dictionary for preservation metadata created by the PREMIS working group. Information on the data dictionary and the working group can be found at: http://www.oclc.org/research/projects/pmwg/, accessed February 2007.
35. The open access principles are formulated in the declarations of Budapest and Berlin. The Budapest Open Access Initiative can be found at: http://www.soros.org/openacess. The Berlin declaration on Open Access to knowledge in the sciences and humanities can be found at: http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html, accessed February 2007.
36. See for instance the Dutch Data Protection Authority (Dutch DPA) in the Netherlands: http://www.dutchdpa.nl, accessed February 2007.
37. VSNU, *The Netherlands code of conduct* (2004).
38. Nestor, "Memorandum on the long-term accessibility" (2006). The memorandum consists of 18 recommendations in the following 4 fields: (1) responsibilities, (2) selection, availability and access, (3) technical measures, and (4) networking and training.
39. The term "Trusted Digital Repository" is introduced and extensively described in RLG working group, *Trusted digital repositories* (2002).
40. http://www.dcc.ac.uk/resource/curation-manual, accessed February 2007.
41. Over 45 instalments are commissioned so far. As of February 2007 the following instalments are available: "Appraisal and Selection", "Preservation Metadata", "Investment in an Intangible Asset", "Curating E-mails", "Archival Metadata", "Metadata" and ""Open Source for Digital Curation".
42. It should be noted that the reports describe the situation in the UK. In the reports references are made to other English speaking countries like the USA and Canada. It seems plausible that the situation described represents the international overall state of art. In some fields other countries have gained specific experiences, e.g. in the application of emulation as preservation strategy (See: Hoeven, van der, "Develepment of" (2005) or in Germany on the certification of trusted repositories (See: Dobratz, "Digitial Repository Certification" (2005).
43. More information on the GDFR and links to references can be found at: http://hul.harvard.edu/gdfr, accessed February 2007.
44. Abrams, "Proceedings of IS&T" (2004).
45. http://www.nationalarchives.gov.uk/pronom, accessed February 2007.
46. The JHOVE software is made available publicly under the GNU General Public License (GPL) from the project website: http://hul.harvard.edu/jhove, accessed February 2007.

47. http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction, accessed February 2007. Adapters have been written for MS Word 2, MS Word 6, Word Perfect, Open Office, MS Works, MS Excel, MS PowerPoint, TIFF, JPEG, WAV, MP3, HTML, PDF, GIF, and BMP.
48. Heery , "Application profiles" (2000).
49. Van Horik describes the Dublin Core registry http://www.dublincore.org/dcregistry/, the Renardus registry http://renardus.sub.uni-goettingen.de/renap/renap.html and the CORES registry http://www.cores-eu.net/registry/, all accessed March 2007. Van Horik, *Permanent Pixels* (2005), 70-72.
50. http://www.alphaworks.ibm.com/tech/uvc, accessed February 2007. The tool, a "proof of concept" provides the long-term access via emulation for JPEG en GIF87a image files. IBM developed the UVC method in cooperation with the Dutch National Library.
51. The website of the TOM project can be found at: http://tom.library.upenn.edu/, accessed February 2007.
52. See for information on solutions for persistent identification: URN: http://www.ietf.org/rfc/rfc2141.txt, "info" URI:53.
    Nestor Working Group Trusted Repositories, "Catalogue of criteria" (2006), 2.
54. Idem, 3.
55. The DRAMBORA toolkit http://www.repositoryaudit.eu/ is an initiative of the "Digital Curation Centre" (DCC) http://www.dcc.ac.uk and the project "Digital Preservation Europe" (DPE) http://www.digitalpreservationeurope.org, all accessed March 2007.
56. Gladney, *Preserving digital information,* (2007)
57. Idem, 254.
58. See Gladney, *Preserving digital information,* (2007) for a complete overview of the components and design of a TDO .
59. http://www.openarchives.org, accessed February 2007.
60. The verbs "Identify", "ListMetaDataFormats" and "ListSets" are related to repositories that are part of a Data Provider. The other three sets are the actual harvesting verbs: they make it possible to locate an individual resource within a repository. When the Service Provider contacts one of the Data Providers for the first time, it will use the "Identify" verb. The Data Provider will respond with a data block, containing relevant information about the organisation hosting the Data Provider and the Data Provider itself. The "ListMetadataFormats" verb yields a list of metadata formats, their validation scheme locations and a metadata prefix.
61. Bekeart, "Augmenting interoperability' (2006).
62. Castelli et all., "Driver architectural specifications" (2006); Kramer, "Possibilities for advanced dissemination" (2006); Hunter, "Scientific publication packages" (2005). By no means it is the intention to cover all system architectures relevant for data curation. Other architectures are e.g. Fedora http://www.fedora.info, and Dspace http://www.dspace.org, accessed March 2007.
63. Kramer,"Possibilities for advanced dissemination" (Berlin, 2006).
64. Electronic Archiving System (EASY): http://easy.dans.knaw.nl, accessed February 2007.
65. AIP stands for "Archival Information Package", a concept part of the OAIS standard. OAIS, *Reference Model* (2003).
66. PANIC stands for "Preservation webservices Architecture for Newmedia and Interactive Collections". See: http://www.metadata.net/panic, accessed February 2007. In Hunter, "Scientific publication packages" (2006), 48, the preservation

of Scientific Information Packages (SPP), as described in paragraph 5.3.2 of this chapter, is implemented with the use of the PANIC system.
67. Hunter, "Semi-automated preservation: (2005), 1.
68. Lyon, ,"Editorial" (2006), 1-2/
69. Beagrie, "E-infrastructure strategy"(2007), 7.
70. Giaretta, "Digital curation and preservation" (2005).
71. Gladney, *Preserving digital information* (2007), 46.

Notes to Long-term preservation for institutional repositories

1. In this context the term "digital curation" is also used. As René van Horik explains in chapter 5, "Data Curation", digital curation not only focusses on the preservation of digital data, but also "relates to the creation of added value and knowledge".
2. Rothenberg, "Ensuring the logevity" (1999). In 1996 the Commission on Preservation and Access and the Research Library Group (RLG) commissioned "Preserving Digital Information. Report of the Task Force On Archiving of Digital Information". This Task Force had the purpose to investigate the means of ensuring "continued access indefinitely into the future of records stored in digital electronic form". See http://www.clir.org/pubs/abstract/pub63.html
3. CCSDS, *Reference model for an Open Archival Information System (OAIS)* (2002), 1-11. The Standard was approved by ISO in 2003 as ISO 14721:2003
4. Jones, *The Preservation Management... handbook*
5. http://portal.unesco.org/ci/en/ev.php-RRL_ID=13366&URL_DO=DO_TOPIC&URL_SECTION=201.html.
6. Mellor, "CAMiLEON" (2003).
7. Entlich, "Digging Up Bits of the Past" (2006).
8. Hockx-Yu, "Digital preservation" (2006), 3.
9. http://www.darenet.nl/en/page/language.view/dare.start
10. www.DSpace.org.
11. www.fedora.info.
12. See for example the CAMiLEON project http://www.si.umich.edu/CAMILEON/, Modular emulation project http://www.kb.nl/hrd/dd/dd_projecten/projecten_emulatie-en.html,, LIFE project http://www.ucl.ac.uk/ls/lifeproject/, ESPIDA project http://www.gla.ac.uk/espida/, PADI subject gateway http://www.nla.gov.au/padi/, Preservation and Long-term Access through networked services http://www.planets-project.eu/, DPE project: Digital Preservation Europe http://www.digitalpreservationeurope.eu/, Caspar project: Cultural, artistic and scientific knowledge for Preservation, Access and Retrieval http://www.casparpreserves.eu/, and the Sherpa DP http://www.sherpadp.org.uk/
13. www.digitalpreservationeurope.eu.
14. www.casparpreserves.eu/.
15. www.planets-project.eu/.
16. See for example http://www.library.cornell.edu/iris/dpworkshop/, or www.kb.nl/hrd/dd/dd_links_en_publicaties/PDF_Guidelines.pdf.
17. James, "Feasibility and requirements study" (2003).
18. Jones, T*he Institutional Repository* (2006), 80.
19. http://droid.sourceforge.net/wiki/index.php/Introduction.
20. http://www.nationalarchives.gov.uk/PRONOM/.

21. http://hul.harvard.edu/gdfr/about.html.
22. Van der Graaf, *The European Repository Lanscape* (2007); Wheatley, *Institutional Repositories* (2004).
23. CCSDS, *Reference model for an Open Archival Information System (OAIS)* (2002). The Standard was approved by ISO in 2003 as ISO 14721:2003
24. CCSDS, *Reference model* (2002), 1-12.
25. CCSDS, *Reference model* (2002), 1-10. See also section 5.3.1. of the chapter on Data curation bij René van Horik.
26. The description of the functional entities in this section is based on chapter 4 of CCDS, Reference model (2002). Several sections of the reference model have been litterally cited.
27. CCSDS, *Reference model* (2002), page 4-1 – 4-2.
28. CCSDS, *Reference model,*(2002), page 2-5.
29. Idem.
30. CCSDS, *Reference model,*(2002), section 3.1
31. Allinson, "OAIS as a reference model"(2006).
32. JISC, "Assessment of UK Data Archive", 81. http://www.jisc.ac.uk/index.cfm?name=project_oais
33. Borghoff, *Long-term preservation of digital documents* (2005*), 136*.
34. For information on Premis see: http://www.loc.gov/standards/premis/.
35. Premis working group, *Data Dictionary* (2005).
36. See: http://www.loc.gov/standards/premis/ The Premis PIG list offers information for repositories planning to implement Premis.
37. Jones, *The Institutional Repository,* 25.
38. http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction.
39. Aschenbrenner, "White Paper on Digital Repoitories" (2005), 28.
40. Digital Preservation Testbed, "Migration" (2001).
41. Mellor, "Migration on request" (2002).
42. Digital Preservation Testbed, "Emulation" (2003).
43. KB / digital preservation testbed.
44. Official statement on emulation for digital preservation, as outcome of the Emulation Expert Meeting 2006 held in The Hague on 20 October 2006. For more information, see: http://www.kb.nl/hrd/dd/dd_projecten/projecten_emulatie-en.html
45. CAMiLEON project, http://www.si.umich.edu/CAMILEON/
46. Modular emulation project of the Koninklijke Bibliotheek and the Nationaal Archief of the Netherlands.
    http://www.kb.nl/hrd/dd/dd_projecten/projecten_emulatie-en.html.
47. http://dioscuri.sourceforge.net/.
48. see http://www.alphaworks.ibm.com/tech/uvc.
49. Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist, http://bibpurl.oclc.org/web/16712.
50. http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf.
51. www.repositoryaudit.eu.
52. Based on the 2005 version, the new version covers the same areas, but the order has changed.
53. RLG/NARA, "Audit checklist for certifying digital repositories", draft for public comment (August 2005) , 32
54. LIFE project, http://www.ucl.ac.uk/ls/lifeproject/.
55. http://www.gla.ac.uk/espida/.
56. http://www.sherpadp.org.uk/.

57. see http://www.nla.gov.au/padi/.
58. D-Lib Magazine http://www.dlib.org, RLG DigiNews http://www.rlg.org/toc.html.

Notes to Appendices

1. Spindler, Rechtliche Rahmenbedingungen (2006).
2. http://eprints.bbk.ac.uk/deposit.hml.
3. http://www.bristol.ac.uk/is/library/collections/rose/rose-licence.html.

References

Abrams, Stephen, and David Seaman, Global digital format registry, "Proceedings of IS&T 2004 Archiving conference", Society for imaging science and technology (San Antonio,Texas, 2004) 83-87.

Allinson, Julie, "OAIS as a reference model for repositories. An evaluation" (UKOLN: Bath, 2006) http://www.ukoln.ac.uk/repositories/publications/oais-evaluation-200607/Drs-OAIS-evaluation-0.5.pdf

Aschenbrenner, Andreas, and Max Kaiser, "White Paper on Digital Repositories" (2005), http://www.uibk.ac.at/reuse/docs/reuse-d11_whitepaper_10.pdf

Barker,Ed, and Charles Duncan, *The Common Information Environment and Creative Commons. Final Report to the Common Information Environment Members of a study on the applicability of Creative Commons Licences*, (Göttingen 2006).

Beagrie, Neil, "Digital curation for science, digital libraries, and individual", *Journal of Digital Curation*, 1/1 (2006), 3-16. http://www.ijdc.net/ijdc/article/view/6/5, accessed 8 February 2007

Beagrie, Neil, "E-infrastructure strategy for research: final report from the OSI preservation and curation working group" (2007). http://www.nesc.ac.uk/documents/OSI/preservation.pdf, accessed 2 March 2007

Bekaert, Jeroen, and Herbert Van De Sompel, "Augmenting Interoperability across scholarly repositories", report of a meeting sponsored and supported by Microsoft, the Andrew W. Mellon Foundation, the Coalition for Networked Information, the Digital Library Federation, and the Joint Information Systems Committee, April 20-21, 2006 (New York, 2006). http://msc.mellon.org/Meetings/Interop/FinalReport, accessed 5 March 2007

Borghoff, Uwe, Peter Rödig, Jan Scheffczyk and Lothar Schmitz, *Long-term preservation of digital documents. Principles and practices* (Springer, 2005).

Bourrion, Daniel, Jean-Louis Boutroy, Claire Giordanengo, and Pascal Krajewski, "Les chercheurs en lettres et sciences humaines et les archives ouvertes, ENSSIB, Mémoire de recherche pour le diplôme de conservateur de bibliothèque, école nacionale supérieure des sciences de l'information et des bibliothèques" (2006) http://hal.archives-ouvertes.fr/docs/00/08/60/84/PDF/chercheurs_LSH_AO_v1.0.pdf, accessd June 2006.

Brogan, Martha, *Context and contributions: building the distributed library* (Washington: Digital Library Federation, 2006). http://www.diglib.org/pubs/dlf106/, accessed 17 February 2007

Callan, Paula, "The development and implementation of a university-wide self-archiving policy at Queensland University of technology (QUT): Insights from the frontline", paper at Institutional Repositories: The Next Stage. Workshop presented at SPARC and SPARC Europe, November 18-19 2004, Washington DC http://eprints.qut.edu.au/archive/00000573/01/callan_sparc.PDF

Carr, Leslie, "Use of navigational tools in a repository", American Scientist Open Access Forum, 9 March 2006 http://users.ecs.soton.ac.uk/harnad/Hypermail/Amsci/5169.html

Carr, Leslie, and Tim Brody, "Size isn't everything: sustainable repositories evidenced by sustainable deposit profiles", *D-Lib Magazine* 13, 7/8 (2007) http://eprints.ecs.soton.ac.uk/13872/

Castelli, Donatella, Paolo Manghi, Pasquale Pagano, Leonardo Candela, Natalia Manola, Vassilis Stoumpos, Friedrich Summann, Marek Imialek and Jaroslaw Wypychowski, "DRIVER Architectural Specifications" (2006) http://www.driver-repository.eu/, accessed 20 February 2007

Chesbrough, Henry, and Richard Rosenbloom, "The role of the business model in capturing value from innovation: evidence from Xerox Corporation's technology spin-off companies", *Industrial and Corporate Change* 11 (3) (2002), 529-555.

Clarke, Roger, "Open source software and open content as models for eBusiness", paper presented at the 17th International eCommerce Conference (Slovenia, 2004) http://www.anu.edu/people/Roger.Clarke/EC/Bled04.html

Consultative Committe for Space Data Systems (CCDS), *Reference model for an Open Archival Information System (OAIS)* (Washington, 2002), http://public.ccsds.org/publications/archive/650x0b1.pdf, accessed November 2007. The Standard was approved by ISO in 2003 as ISO 14721:2003

Crews, Kenneth, and Gerard van Westrienen, "Copyright, publishing and scholarship.The 'Zwolle Group' Initiative for the Advancement of Higher Education", *D-Lib Magazine* 13, 1/2 (2007).

Crow, Raym, "The Case for Institutional Repositories: A SPARC Position Paper", *ARL Bi-monthly Report 223*. (2002) http://www.arl.org/sparc/bm~doc/ir_final_release_102.pdf

Davis, Philip, and Matthew Connelly, "Institutional Repositories, Evaluating the Reasons for Non-use of Cornell University's Installation of Dspace", *D-Lib Magazine*, 13, 3/4 (2007) http://www.dlib.org/dlib/march07/davis/03davis.html

Day, Michael, *Institutional repositories and research assessment* (UKOLN: Bath, 2004) http://eprints-uk.rdn.ac.uk/project/docs/studies/rae/rae-study.pdf

Digital Preservation Testbed, "Emulation: context and current status" (The Hague, 2003) http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf

Digital Preservation Testbed, "Migration: Context and Current Status", White Paper (December 2001), http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=185&lang=en

Dobratz, Susanne, and Astrid Schoger, Digital Repository Certification: a Report from Germany', *RLG DigiNews* 9 (5) (2005) http://www.rlg.org/en/page.php?Page_ID=20793#article3, accessed 8 February 2007

Domingus, Marlon, "Research unleashed?", paper for the conference Bridging the North-South Divide in Scholarly Communication on Africa. Threats and Opportunities in the Digital Era, 6-8 Sept. 2006 (Leiden, 2006) http://www.ascleiden.nl/Pdf/elecpublconfdomingus.pdf

Ducloy, Jacques et al., (2006) "Metadata towards an e-research cyberinfrastructure. The case of francophone PhD theses." Paper for the 2006 Annual Dublin Core Conference, http://artist.inist.fr/IMG/pdf/MetadataV14_4_.pdf

Emulation Expert Meeting, "Official statement on emulation for digital preservation, as outcome of the Emulation Expert Meeting 2006 held in The Hague on 20 October 2006" (internet, 2006) http://www.kb.nl/hrd/dd/dd_projecten/projecten_emulatie-en.html

Entlich, Richard, and Ellie Buckley, "Digging Up Bits of the Past: Hands-on With Obsolescence", *RLG Digi News*, 10/5 (2006) http://www.rlg.org/en/page.php?Page_ID=20987#article1

Commission of the European Communities, "Green Paper on Copyright and the Challenge of Technology – copyright issues requiering immediate action" *COM* (88) 172 final (Brussels, 1988).

European Parliament, "Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases", *Official Journal* L 077 (1996), 20–28.

Feijen, Martin, and Annemiek van der Kuil, "A Recipe for Cream of Science: Special Content Recruitment for Dutch Institutional Repositories", *Ariadne,* 45 (2005) http://www.ariadne.ac.uk/issue45/vanderkuil

Fox, Edward, "The 5S Framework for Digital Libraries and Two Case Studies: NDLTD and CSTC" (Blacksburg,1999) http://docs.ndltd.org:8080/dspace/handle/2340/32

Giaretta, David, et al., "Digital curation and preservation: Defining the research agenda for the next decade", report from the Warwick Workshop – 7 & 8 November 2005 (2005) http://www.dcc.ac.uk/events/warwick_2005/Warwick_Workshop_report.pdf, accessed 2 March 2007

Gierveld, Heleen, "Considering a Marketing and Communications Approach for an Institutional Repository", *Ariadne,* 49 (2006) http://www.ariadne.ac.uk/issue49/gierveld/

Gladney, Henry, *Preserving digital information* (Springer: Heidelberg, 2007)

Graaf, Maurits van der, and Kwame van Eijndhoven, *The European Repository Landscape . Inventory study into the present level of OAI compliant digital repository activities in the EU (AUP: Amsterdam, 2007) www.driver-community.eu .*

Harnad, Stevan, "Publish or Perish – Self-Archive to Flourish: The Green Route to Open Access", *ERCIM News* 64 (2006) http://eprints.ecs.soton.ac.uk/11715/

Harnad, Stevan, "Self-archiving should be mandatory", *Research Information* (2006) [website] http://eprints.ecs.soton.ac.uk/12738/

Harnad, Stevan, "The Golden and Green Roads to Open Access", [AmSci Open Access Forum, posted November 14, 2003]

Harnad, Stevan, Tim Brody, François Vallieres, Leslie Carr, Steve Hitchcock, Yves Gingras, Charles Oppenheim, Heinrich Stamerjohanns, and Eberhardt Hilf, "The Access/Impact Problem and the Green and Gold Roads to Open Access", *Serials review* 30/4 (2004) http://eprints.ecs.soton.ac.uk/10209/

Heery, Rachel, and Sheila Anderson, "Digital Repositories Review", UKOLN and AHDS Report, (UKOLN: Bath, 2005) http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf

Heery, Rachel, and Manjula Patel, "Application profiles: mixing and matching metadata schemas", *Ariadne,* 5 (2000) http://www.ariadne.ac.uk/issue25/app-profiles/, accessed 21 February 2007

Hess, Thomas, Rolf Wigand, Florian Mann, and Benedikt von Walter, *Open access and Science publishing, Results of a study on Researchers acceptance and use of Open access publishing* (Munich, 2007), 10.

Hey, Jessie, "Targeting academic research with Southampton's institutional repository", *Ariadne* 40 (2004) http://www.ariadne.ac.uk/issue40/hey/

Hill, Linda., Scott Crosier, Terence Smith, and Michael Goodchild, "A content standard for computational models", *Dlib Magazine,* 7/6 (2001) http://www.dlib.org/dlib/june01/hill/06hill.html, accessed 12 February 2007.

Hirtle, Peter, "Author Addenda. An examination of five alternatives", *D-Lib Magazine* 12 /11 (2006).

Hockx-Yu, Helen, "Digital preservation in the context of institutional repositories", *E-LIS – E-prints in Library and Information Science* (online database) (2006) http://eprints.rclis.org/archive/00007351/

Hoeven, Jeffrey van der, Raymond van Diessen, and Kornelis van der Meer, "Development of a Universal Virtual Computer (UVC) for long-term preservation of digital objects", *Journal of Information Science*, 31/3 (2005), 196-208

Horik, René van, *Permanent pixels. Building blocks for the longevity of digital surrogates of historical photographs* (DANS: The Hague, 2005).

Houghton John, Colin Steele, and Peter Sheehan, *Research communication costs in Australia: Emerging opportunities and benefits*, a report to the Department of Education, Science and Training. (Melbourne, 2006) http://www.dest.gov.au/NR/rdonlyres/0ACB271F-EA7D-4FAF-B3F7-0381F441B175/13935/DEST_Research_-Communications_Cost_Report_Sept2006.pdf

Hunter, Jane, "Scientific Publication Packages – A selective approach to the communication and archival of scientific output", *Journal of Digital Curation* 1/1 (2006) 3-16. http://www.ijdc.net/ijdc/article/view/8/7, accessed 8 February 2007

Hunter, Jane, and Sharmin Choudhury, "Semi-Automated Preservation and Archival of Scientific Data using Semantic Grid Services", Semantic Infrastructure for Grid Computing Applications Workshop at the International Symposium on Cluster Computing and the Grid, CCGrid 2005. (Cardiff, 2005) http://www.metadata.net/panic/papers/SIGAW2005_paper.pdf, accessed 21 February 2007

IFLA Study Group on the Functional Requirements for Bibliographical Records, *Functional requirements for Bibliographic records – Final Report* (K.G. Saur: Munich, 1998) http://www.ifla.org/VII/s13/frbr/frbr.pdf, accessed 15 January 2007

Jacobs, Neil, (ed.), *Open Access: Key Strategic, Technical and Economic Aspects* (Chandos Publishing: Oxford, 2006)

James, Hamisj, Raivo Ruusalepp, Sheila Anderson and Stephen Pinfield, "Feasibility and requirements study on preservation of e-prints", report commissioned by the Joint Information Systems Committee (London, 2003) http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf

JISC, "Assessment of UK Data Archive and The National Archives compliance with OAIS/METS"[website] http://www.jisc.ac.uk/index.cfm?name=project_oais

Jones, Maggie, and Neil Beagrie / The Digital Preservation Coalition, *The Preservation Management of Digital Material Handbook* [online text] http://www.dpconline.org/graphics/handbook/

Jones, Richard, Theo Andrew, and John MacColl, *The Institutional Repository* (Chandos Publishing: Oxford, 2006)

Kaczmarek, Joanne, Patricia Hswe, Janet Eke, and Thomas Habing, "Using the Audit Checklist for the Certification of a Trusted Digital Repository as a Framework for Evaluating Repository Software Applications. A Progress Report", *D-Lib Magazine*, 12/12 (2006) http://www.dlib.org/dlib/december06/kaczmarek/12kaczmarek.html, accessed 12 February 2007

Korn, Naomi, and Charles Oppenheim, "Creative Commons Licences in Higher and Further Education: Do we care?" *Ariadne* 49 (2006).

Kramer, Rutger, "Possibilities for Advanced Dissemination and Durable Storage of Scientific Data on the Grid", 2nd International Conference on Trends in Enterprise Application Architectures (Berlin, 2006)

Lagoze, Carl, and Jane Hunter, "The ABC ontology and model", *Jounal of digital information* 2/2 (2001) http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/, accessed February 2007

Lord, Philip, and Alison Macdonald, "e-Science Curation Report. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision. (Twickenham, 2003) http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf accessed February 2007

Lucas, André, and Henri-Jacques Lucas, *Traité de la propriétaire littéraire et artistique*, (Litec : Paris, 2006)
Lynch, Clifford, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age", *ARL: A Bimonthly Report* 226, (2003) http://www.arl.org/resources/pubs/br/br226/br226ir.shtml
Lyon, Liz, "Dealing with data: roles, rights, responsibilities and relationships: a consultancy report" (UKOLN: Bath, 2007) http://www.jisc.ac.uk/media/documents/programmes/digital_repositories/dealing_with_data_report-final.pdf
Lyon, Liz, and. and Chris Rusbridge, "Editorial", *Journal of Digital Curation,* (1/1, 2006), 1-2, http://ijdc.net/ijdc/article/view/16/24
Mackie, Morag, "Filling Institutional Repositories: Practical strategies from the DAEDALUS Project", *Ariadne* 39 (2004) http://www.ariadne.ac.uk/issue39/mackie/
Magron, Agnès, "Auto-archivage des publications scientifiques : Synthèse d'enquêtes menées auprès des chercheurs, Instiut des Sciences de l'Homme" (Lyon, 2007) http://archivesic.ccsd.cnrs.fr/docs/00/15/15/75/PDF/EnquetespratiquesOA.pdf
Markey, Karen, Soo Young Rieh, Beth St. Jean, Jihyun Kim, and Elizabeth Yakel, *Census of Institutional Repositories in the United States: MIRACLE Project Research Findings* (web only publication, 2007) http://www.clir.org/pubs/abstract/pub140abst.html
Mellor, Phil, "CAMiLEON: Emulation and BBC Domesday", *RLG Diginews*, 7/2 (2003), http://www.rlg.org/legacy/preserv/diginews/v7_n2_feature3.html
Mellor, Phil, Paul Wheatley, and Derek Sergeant, "Migration on Request: A practical technique for digital preservation" (Leeds, 2002) http://www.si.umich.edu/CAMILEON/reports/reports.html
Migus, Arnold, "Libre accès à l'information scientifique et technique : Actualités, problématiques et perspectives", (website, 2006) http://openaccess.inist.fr/article.php3?id_article=127
Moffat Malcom, "'Marketing' with metadata – how metadata can increase exposure and visibility of online content"(website, 2006) http://www.icbl.hw.ac.uk/perx/advocacy/exposingmetadata.htm
Mossink,Wilma, *Auteursrechten op wetenschappelijke publicaties* (SURF//Open Universiteit, 1999).
Mossink, Wilma, and Ralph Weedon, Report on institutional copyright policies in the Netherlands & the UK (2006).
Nestor, "Nestor Memorandum on the long-term accessibility of digital information in Germany"(2006) http://www.langzeitarchivierung.de/downloads/memo2006-e.pdf, accessed 26 February 2007
Nestor Working Group Trusted Repositories – Certification, "Catalogue of criteria for trusted digital repositories", *Nestor-materials 8* (Franfurt am Main, 2006) http://edoc.hu-berlin.de/series/nestor-materialien/8/PDF/8.pdf, accessed March 2007
Nixon, William, "The evolution of an institutional e-prints archive at the University of Glasgow" *Ariadne* 32 (2002) http://www.ariadne.ac.uk/issue32/eprint-archives/intro.html
OECD, *Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 – Final Communique,* http://www.oecd.org/document/15/0,3343,en_2649_201185_25998799_1_1_1_1,00.html.

Pappenberger, Karlheinz, "Strategien zur Umsetzung von Open Access an der UB Konstanz", paper for the German Librarian Conference in Dresden, 22.03.2006 – Themenkreis 8.03. Bibilotheek der Universitaet Konstanz (2006) http://www.opus-bayern.de/bib-info/volltexte/2006/202/pdf/pappenberger_dresden2006.pdf

Pickton, Margaret, and Cliff McKnight., "Research students and the Loughborough institutional repository", *Journal of Librarianship and Information Science,* 38/4 (2006), http://lis.sagepub.com/cgi/reprint/38/4/203

Pinfield, Stephen, Mike Gardner, and John MacColl, "Setting up an institutional e-print archive" *Ariadne* 31 (2002) http://www.ariadne.ac.uk/issue31/eprint-archives/intro.html

Porter, George, "Tips for filling an institutional repository", *Open Access News* (Internet, 2006) http://www.earlham.edu/~peters/fos/2006_08_13_fosblogarchive.html

Premis working group, *Data Dictionary for Preservation Metadata Final Report* (Dublin USA, 2005) http://www.oclc.org/research/projects/pmwg/premis-final.pdf

Rappa, Michael, "Business models on the Web: managing the digital enterprise"[website]. (North Carolina State University, USA) http://digitalenterprise.org/models/models.html

RLG/NARA, "Audit checklist for certifying digital repositories", draft for public comment (August 2005)

RLG/OCLC Working Group on Digital Archive Attributes, *Trusted Digital Repositories: Attributes and Responsibilitie. An RLG-OCLC Report* (Mountain View, CA, 2002) http://www.rlg.org/longterm/repositories.pdf, accessed February 2007

Roosendaal, Hans, Peter Geurts, and Paul van der Vet, "Developments in scientific communication: considerations on the value chain" *Information Services and Use* 21 (2001), 13-32

Roosendaal, Hans, and Peter Geurts "Forces and functions in scientific communication: an analysis of their interplay" *CRISP* 97 (1998).

Ross, Seamus, and Andrew McHugh, "Audit and Certification of Digital Repositories : Creating a Mandate for the Digital Curation Centre (DCC)", *RLG DigiNews,* 9/5 (2005) http://www.rlg.org/en/page.php?Page_ID=20793#article1, accessed 8 February 2007

Rothenberg, Jeff, "Ensuring the longevity of digital information", (RAND: Santa Monica, 1999) http://www.clir.org/pubs/archives/ensuring.pdf

Rowlands, Ian, and Dave Nicholas, "The changing scholarly communication landscape: an international survey of senior researchers", *Learned Publishing,* 19/1 (2006), 31-55, http://www.publishing.ucl.ac.uk/papers/2006Rowlands_Nicholas.pdf

Sale, Arthur, "Comparison of content policies for institutional repositories in Australia", *First Monday,* 11/4 (2006) http://www.firstmonday.org/issues/issue11_4/sale/

Sale, Arthur, "The acquisition of open access research articles", *First Monday,* 11/10 (2006) http://www.firstmonday.org/issues/issue11_10/sale/index.html

Sale, Arthur, "The Patchwork Mandate", *D-Lib Magazine* 13 number 1/2 (2007) http://www.dlib.org/dlib/january07/sale/01sale.html

Sparks, Sue, *JISC's Disciplinary Differences Report* (London, 2005) http://www.jisc.ac.uk/media/documents/themes/infoenvironment/disciplinarydifferencesneeds.pdf

Spindler Gerald, *Rechtliche Rahmenbedingungen von Open access-Publikationen*, (Goettingen, 2006).

Suber, Peter, "Open Access Overview – Focusing on Open Access to peer-reviewed research articles and their preprints" [website] http://www.earlham.edu/~peters/fos/overview.htm, accessed 9 August 2007

Sufi, Shoaib, and Brian Mathews, "CCLRC Scientific Metadata Model: Version 2", CCLRC Technical Report DL-TR-2004-001 (2004), http://epubs.cclrc.ac.uk/bitstream/485/csmdm.version-2.pdf, accessed 12 February 2007]

Svenonius, Elaine, *The intellectual foundation of information organization* (MIT Press: Cambridge, 2000)

Swan, Alma, "The culture of open access: researchers' views and responses", Neil Jacobs (ed.), *Open Access: Key Strategic, Technical and Economic Aspects* (Chandos Publishing: Oxford, 2006) 52-59. http://eprints.ecs.soton.ac.uk/12428/01/asj7.pdf

Swan, Alma, and Chris Awre, *Linking UK repositories: Technical and organisational models to support user-oriented services across institutional and other digital repositories: Scoping study report* (2006) http://www.jisc.ac.uk/uploaded_documents/Linking_UK_repositories_report.pdf Appendix: http://www.jisc.ac.uk/uploaded_documents/Linking_UK_repositories_appendix.pdf

Swan, Alma, and Sheridan Brown, *Open Access Self-archiving : An Author Study* (Key Perspectives Ltd. JISC Study, 2005) http://eprints.ecs.soton.ac.uk/10999/01/jisc2.pdf

Swan, Alma, and Sheridan Brown, *Researcher awareness and access to open access content through libraries: A study for the JISC Scholarly Communications Group* (Key Perspectives Ltd.: Truro, 2007) http://eprints.ecs.soton.ac.uk/14412/

Swan, Alma, Paul Needham, Steve Probets, Adrienne Muir, Ann O'Brien, Charles Oppenheim, Rachel Hardy, and Fytton Rowland, "Delivery, management and access model for E-prints and open access journals within further and higher education", Report of a JISC study (2004) http://eprints.ecs.soton.ac.uk/11001/

Timmers, Paul, "Business models for electronic markets", Yves Gadient, Beat Schmid and Dorian Selz, *EM-Electronic Commerce in Europe. EM- Electronic Markets* 8/2 (1998) www.electronicmarkets.org/modules/pub/view.php/electronicmarkets-183

Tindemans, Peter, "Key stakeholders pledge to a strategic approach to preserve the digital records of science", *Journal of Digital Curation* (1/1) 2006, 83-89. http://www.ijdc.net/ijdc/article/view/13/3, cited 2 February 2007

University of Houston Institutional Repositories Task Force, *SPEC Kit No. 292: Institutional Repositories* (Association of Research Libraries: Washington, 2006) http://www.arl.org/resources/pubs/spec/complete.shtml

VSNU, *The Netherlands Code of Conduct for Scientific Practice. Principles of good scientific teaching and research* (25 October 2004), http://www.vsnu.nl/web/show/id=87061/langid=42, accessed February 2007.

Waaijers, Leo, "From libraries to libratories", *First Monday,* 10/12 (2005) http://www.firstmonday.org/issues/issue10_12/waaijers/index.html

Waller, Martin and Robert Sharpe, *Mind the gap: assessing digital preservation needs in the UK* (Digital preservation coalition, 2006) http://www.dpconline.org/docs/reports/uknamindthegap.pdf, accessed 2 March 2007

Wheatley, Paul, *Institutional Repositories in the Context of Digital Preservation*, DPC Technology Watch Series Report 04-02, (Digital Preservation Coalition: York, 2004) http://www.dpconline.org/docs/DPCTWf4word.pdf

Wojciechowska, Anna, "Analyse d'usage des archives ouvertes dans le domaine ddes mathématiques et l'informatique" (2006) http://archivesic.ccsd.cnrs.fr/action/

open_file.php?url=http://archivesic.ccsd.cnrs.fr/docs/00/06/27/22/PDF/sic_00001735.pdf&docid=62722

Index