LEAGUE OF EUROPEAN RESEARCH UNIVERSITIES

ADVICE PAPER
No.14 - December 2013

LERU Roadmap for Research Data

LERU Research Data Working Group

LEAGUE OF EUROPEAN RESEARCH UNIVERSITIES

University of Amsterdam - Universitat de Barcelona - University of Cambridge - University of Edinburgh - University of Freiburg - Université de Genève - Universität Heidelberg - University of Helsinki - Universiteit Leiden - KU Leuven - Imperial College London - University College London - Lund University - University of Milan - Ludwig-Maximilians-Universität München - University of Oxford - Pierre & Marie Curie University - Université Paris-Sud - University of Strasbourg - Utrecht University - University of Zurich

LEAGUE OF EUROPEAN RESEARCH UNIVERSITIES

Contents

Contributors

Pablo Achard (University of Geneva)
Paul Ayris, UCL (University College London)
Serge Fdida (UPMC, Paris)
Stefan Gradmann (University of Leuven)
Wolfram Horstmann (University of Oxford)
Ignasi Labastida (University of Barcelona)
Liz Lyon (University of Bath)
Katrien Maes (LERU)
Susan Reilly (LIBER)
Anja Smit (University of Utrecht)

INTRODUCTION

The LERU Roadmap for Research Data represents a recognition that LERU universities now work in an era of data-driven science. The Royal Society report Science as an Open Enterprise,[1] co-ordinated by Professor Geoffrey Boulton from the University of Edinburgh, has set the tone of debate and the direction of travel for LERU members.

It is also important to note the linkage between Research Data Policy, Technology and Support. To promote conscious and successful use of research data, these three aspects should be offered simultaneously to researchers. Projects that merely focus on one or two of these aspects are doomed to fail, as well as projects where policy, support and services are not aligned. A co-ordinated and parallel approach is therefore crucial.

This Advice Paper was requested to be written by the LERU Rectors in the realisation that research data, and the prospect of open data, is an issue on which LERU universities need to take a position. Opportunities for alleviating societal problems can be enhanced by researchers sharing their data.[2] It is therefore obvious that LERU members need to act. In 2011, the LERU community of Chief Information Officers produced a Roadmap for Open Access to publications[3] ratified by the LERU Rectors. Now, the CIO Community has produced a second Roadmap, this time for research data.

This Roadmap looks at the challenges posed by research data in seven chapters, which concentrate on issues such as policy, leadership, research data infrastructure, costs, advocacy, description and legal issues, skills, roles and responsibilities. Through selected case studies and examples from LERU universities, it is possible to see how individual LERU members are tackling these challenging issues.

The resulting Roadmap, like its predecessor on Open Access to research publications, presents a series of blueprints which LERU members, indeed any European university, could use to begin to tackle the challenges which research data poses. It also has a series of messages for researchers, research institutions, support services and policymakers.

Paul Ayris (UCL)
on behalf of the LERU Research Data Working Group

December 2013

1 See http://royalsociety.org/policy/projects/science-public-enterprise/report/
2 For a case study on the Spanish E Coli strain, see n. 1
3 See http://www.leru.org/files/publications/LERU_AP8_Open_Access.pdf

Executive Summary

The LERU Roadmap for Research Data plots a course which LERU members can choose to follow to implement sound research data management practices at institutional level. The Roadmap is divided into six chapters, with the seventh being devoted to a series of Recommendations which stem from the text.

Chapter 1 looks at the ideas of Policy and Leadership in this field. It shows that universities have responded to a greater or lesser degree to data policy directives. It argues that what is needed are institutional data management policies and accompanying Roadmaps for Research Data management.

Chapter 2 looks at the issue of Advocacy, which the Roadmap identifies as crucial to successful data sharing. The Roadmap identifies incentives and barriers to data sharing, along with suggestions for how to overcome the reluctance of researchers to share in this way. Open research data is advocated as a goal for all researchers, where this is possible. This requires leadership at an institutional level. University support services are well placed to advocate for best practice in research data management and data citation. Advocacy can underline the rewards inherent in data sharing, help to make data visibile, increase collaboration and data reuse, and help to build the necessary trust to make all this happen.

Chapter 3 looks at a range of issues involved in the management of research data: Selection and Collection, Curation, Description, Citation and Legal Issues. For selection and curation, the Roadmap takes as its starting point the ODE Data Publication Pyramid and recommends that the LERU research community should undertake further work to identify which of the strata of research data identified by the pyramid can be made available for sharing and re-use, and which can be open. In terms of Data Curation, the Roadmap first analyses the research workflow and then suggests how the necessary infrastructures can be created. For Description, the Roadmap underlines the difficulties inherent in encourging researchers accurately to describe their data. For Citation, examples of best practice in data citation are provided. The final section, on Legal Issues, analyses the European copyright framework and suggests that a Fair Dealing Exception is required to enable Text and Data Mining tools and techniques to flourish in an era of data-driven science.

Chapter 4 looks at Research Data infrastructure. These infrastructures can be classified into four types: research data itself, data management tools, technical components and staffing. Research data infrastructure needs to offer a generic framework to accommodate the wide variety of research activities which will make use of it. An overview of research data management tools is provided and the chapter highlights that the 'long tail' of research data residing on local desktops, hard disks and servers might well comprise a bigger challenge than 'big data'. In terms of technical components, the chapter outlines how these components are distributed across the university. For staffing, the chapter likewise identifies that this resource is distributed across the institution, and that ideally it should be organised as a coherent support service.

Chapter 5 tackles the difficult issue of Costs. There is no one single model which can be used to calculate costs. It provides two Case Studies, for the University of Oxford and UCL (University College London) to give indicative costs for service provision. The chapter shows that Cost Benefits can sometimes provide a framework for judging the cost effectiveness of research data curation. It also shows who is likely to meet the costs – research funder, national collaborative service, or the university itself.

Chapter 6 looks at Roles, Responsibilities and Skills. The chapter undertakes an analysis of the different roles needed/involved in research data management and the responsibilities that these postholders have. It suggests that a new concept of Data Scientist has the potential to become a new role in its own right. The chapter also identifies the training requirements needed of a range of participants such as postgraduates/PhD students, senior researchers, librarians and data scientists.

The final Chapter, Chapter 7, brings together 44 Recommendations drawn from the Roadmap and allocates them to specific audiences: institutional policy and decision makers in LERU and other universities, all those involved in the curation of research data, researchers and their institutions, LERU members and the LERU community of Chief Information Officers, and the bodies of the European Union.

Setting the context

1. This Roadmap looks at the challenges posed by research data management (RDM) from six viewpoints:
- Policy and Leadership
- Advocacy
- Selection and Collection, Curation, Description, Citation, Legal Issues
- Research data Infrastructure
- Costs
- Roles, Responsibilities and Skills

2. Research data,[4] from the point of view of the institution with a responsibility for managing the data, includes:
- all data which is created by researchers in the course of their work, and for which the institution has a curatorial responsibility for at least as long as the code and relevant archives/record keeping acts require, and
- third-party data which may have originated within the institution or come from elsewhere.

3. Research institutions already manage different kinds of data. It is, therefore, possible to consider a definition of research data to some extent in terms of what it is not:
- administrative data consists of records of payrolls, student enrolments, research assessment, and so on. Some administrative data relates to research projects and may need to be treated as research data. However, for the most part it is treated independently within the institution in terms of data management policies, procedures and strategies
- teaching data comprises courseware and other resources which are part of the teaching function of a university. Again, this may be of interest to a research project, but it is usually managed independently
- research publications can be regarded as data, but for the most part these are well taken care of outside the institution, by publishers and the like. Even when held within the institution, either on open access or for research reporting purposes, these tend to be managed separately from other research data.

4. "A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike."[5] Open data is, therefore, the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.[6] The general position taken by this Roadmap is that data should be made open at the most appropriate time, when the researcher/research group is ready to make this material available. However, as the Case Study in Chapter 1 illustrates, not all data can be open. There may be third party contractual agreements, especially with commercial funders, which mean that data is governed by a Partnership agreement with provisions for how the resulting research data is managed. There may be legislative obligations, for example data protection legislation, which mean that data cannot be open in order to protect the identities and information concerning individuals. Researchers may wish to stipulate provisions regarding disclosure of data, not wishing to make it open until they have completed their cycle of publications from the data that they have collected.

5. The Royal Society Report Science as an Open Enterprise illustrates the benefits that accrue to research, and to Society, by researchers being willing to make their research data available for sharing and re-use. There is therefore a link between the management of research data and research integrity. The benefits of sound research data management are reflected in this Roadmap.

6. Good data management is imperative and should be underpinned by interoperable research information management practices. This enables visibility and sharing where appropriate, as well as storage, preservation, reporting and business intelligence.

7. In terms of Policy and Leadership, the Roadmap underlines the need for each LERU institution to construct its own Roadmap to help guide the de-

4 Taken from http://ands.org.au/guides/what-is-research-data.html

5 See http://opendefinition.org/

6 Auer, S. R.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. (2007). "DBpedia: A Nucleus for a Web of Open Data". *The Semantic Web*. Lecture Notes in Computer Science 4825. p. 722. http://dx.doi.org/10.1007/978-3-540-76298-0_52

velopment of policy, to be clear about roles and responsibilities and to act as a framework for implementation of RDM activities in LERU institutions.

8. Advocacy is identified as a crucial element for LERU members. Data-driven science needs to be subject-oriented. LERU universities need to foster a debate amongst stakeholders and disciplines around research data management and data sharing; to develop and clearly articulate incentives for researchers to make their data open; and to promote best practice in data management, citation and interoperability to increase awareness of the importance of data management itself (which is lacking in some areas), to increase the visibility of research data and also for audit and reporting.

9. In the chapter on Selection and Collection, Curation, Description, Citation and Legal Issues, the Roadmap analyses a number of issues which need to be addressed by those involved in research data management and by researchers. A range of individual Recommendations addresses the whole range of issues covered by the chapter. Practical support for researchers should be organized. It is recommended that institutions and researchers work together to identify best practice when citing data. Publication of 'basic principles', general guidelines and examples to help researchers in how to cite data are very common and helpful. This information should include links to existing demands from funding agencies, publishers and data centres, specific to the different disciplines.

10 The chapter on Research Data Infrastructure concludes that this is an essential prerequisite for today's and tomorrow's research. The renewed importance of research data, especially its volume, moves demands on university facilities into the focus – also because funders expect researchers to make their data openly available in the long-term.

11. The chapter on Costs addresses the issue by looking at two Case Studies, the universities of Oxford and UCL (University College London). There is a remarkable similarity in the range of costs which the institutions have individually identified. The CIO community in LERU has an important role to play in collecting and sharing information on costing for research data management and could act as a focus of best practice in this area.

12. The final chapter, on Roles, Responsibilities and Skills, looks at the issue of how the necessary skills and knowledge in research data management are fostered and embedded institutionally. A whole range of Recommendations addresses these issues, for example: data management courses should be embedded within postgraduate training; LERU institutions should introduce specific job profiles with career paths for data preparation and quality assurance staff – such staff may be embedded in research groups or hosted in data centres or libraries.

13. The final chapter brings all the Recommendations in the LERU Roadmap for Research Data together, and addresses them to the most appropriate audience:

- Institutional policy and decision makers at LERU and other universities
- Those involved in the curation of research data
- Researchers and their institutions
- LERU Community of Chief Information Officers
- Bodies of the EU

Chapter 1: Policy and Leadership

Funder Policy Context

14. During the past few years there has been a growing international momentum behind open data developments and the effective curation and management of research data to support use and re-use of data outputs. In 2010, the European Union explicitly stated "*publicly funded research should be widely disseminated through Open Access publication of scientific data and papers*"[7] and sought to ensure "*dissemination, transfer and use of research results, including through open access to publications and data from publicly funded research*".[8] In 2012, the Commission announced that it will "*provide a framework and encourage open access to research data in Horizon 2020*"[9]. As an example, in Germany, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) recently approved the paper "*Taking Digital Transformation to the Next Level*"[10] and funded a programme for research data infrastructures[11]. In the UK in 2012, the influential Royal Society Report[12] made a series of Recommendations to funders and institutions to promote open data. UK research funding agencies have collaborated to publish the RCUK Common Principles on Data Policy[13] and research councils have published their individual data policy directives.

15. In the United States in May 2013, the Obama Administration published a ground-breaking Open Data policy[14] which covers data and "*requires agencies to collect or create information in a way that supports downstream information processing and dissemination activities.*" In June 2013, the G8 leaders signed the G8 Open Data Charter[15] which included five over-arching principles to support open data and innovation.

16. All of these policy initiatives have greatly strengthened the global imperative to curate, manage and share research data as part of the wider open data environment. The reports and statements also have implications for capacity-building within the sector, as well as policy compliance, potential contractual requirements and emerging professional good practice.

Institutions and Stakeholder Engagement

17. Within higher education, research-intensive universities have responded to a greater or lesser degree to these data policy directives. Many institutional stakeholders are involved in the research data lifecycle and have a role in the creation, collection, processing, analysis, curation and preservation of research data[16]. These include (but are not limited to) researchers, academic faculty, senior managers such as Vice-Rectors Research, doctoral training centres, planning offices, research support staff, legal offices, IT services, libraries and information services. External parties such as disciplinary data centres and publishers are also critical elements in the data lifecycle. At a disciplinary level, there are examples where research data management (RDM) solutions have been implemented. Some universities have also taken the lead in promoting RDM and in developing essential human and technical infrastructure support services. In some cases, a cross-institutional working group or board has been created to oversee research data management activities. This can be a highly effective way to bring together key institutional players and progress planning and implementation.

7 Digital Agenda for Europe, 2.5.2, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0245:FIN:EN:HTML
8 Europe 2020 Flagship Initiative, http://ec.europa.eu/research/innovation-union/pdf/innovation-union-communication_en.pdf
9 Towards better access to scientific information: Boosting the benefits of public investments in research, http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf
10 Taking Digital Transformation to the Next Level, http://www.dfg.de/en/service/press/press_releases/2012/press_release_no_29/index.html
11 Handling Big Data and Small Data in a Sustainable Way, http://www.dfg.de/en/service/press/press_releases/2013/press_release_no_06/index.html
12 Royal Society Report Science as an Open Enterprise (June 2012) http://royalsociety.org/policy/projects/science-public-enterprise/report/
13 RCUK Common Principles on Data Policy http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx
14 http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf
15 G8 Open Data Charter https://www.gov.uk/government/publications/open-data-charter
16 Liz Lyon, 'Informatics Transform: re-engineering libraries for the data decade' (2012) in the International Journal of Digital Curation, vol. 7, no. 1, pp. 126-138; available at http://www.ijdc.net/index.php/ijdc/article/view/210

Roadmaps

18. Libraries, Information Services and Research Services have also demonstrated leadership in producing Roadmaps, in particular in response to the UK Engineering and Physical Sciences Research Council (EPSRC) requirement, which placed the responsibility on institutions, rather than individual research grant recipients. The concept of a Roadmap has been interpreted in different ways by institutions; exemplars have been collected by the Digital Curation Centre.[17] One such is from the University of Bath[18] which describes how the institution will respond to the nine EPSRC expectations. This Roadmap was approved by the Vice-Chancellor's Group and the tasks are now being implemented by a range of stakeholders across the university. Implementation of the Roadmap actions is central to the successful embedding of good research data management practice. This work was initiated by a Jisc-funded innovation programme project (Research360) and is now being taken forward by the University Library with additional institutional funding support for two new data support roles. Similar approaches are being adopted at other UK institutions, such as the University of Oxford and UCL.

Institutional Policy Development

19. One of the key elements of these Roadmaps is the development and adoption of an institutional data policy with accompanying guidance. Data policy exemplars have also been collected by the DCC[19] and range from the aspirational to the more pragmatic approaches which can be implemented at local level. In many cases, the data policies have been approved by university research committees and cover the roles and responsibilities of institutional stakeholders. A case study of data policy development is provided from UCL.

Conclusions and Recommendations

20. Individual LERU members should explore the formation of an RDM Steering Group or similar which brings together the range of critical institutional stakeholders and provides a forum for planning and operational oversight.

21. Each LERU member should consider developing an institutional Roadmap for Research Data (if they have not done so already) which sets out the strategic objectives, tasks and actions required for compliance with research funder directives.

22. Every LERU member should develop and promulgate an institutional data policy which clarifies institutional roles and responsibilities for RDM to all stakeholders in the RDM process.

17 EPSRC Roadmaps. http://www.dcc.ac.uk/resources/policy-and-legal/epsrc-institutional-roadmaps

18 The University of Bath Roadmap for EPSRC is available at http://blogs.bath.ac.uk/research360/2012/06/the-university-of-bath-roadmap-for-epsrc/

19 Institutional data policies http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies/uk-institutional-data-policies

Case Study 1 - UCL Research Data Policy

In Summer 2013, UCL (University College London) signed off its Research Data policy.[20] The policy was generated by a number of drivers for UCL researchers:

- Policy requirements by external funders
- Need to inform UCL researchers of their roles and obligations in an era of data-driven research
- Policy creation to raise awareness of fundamental research issues facing UCL

The UCL policy, which is given below, identifies who is responsible for data management. Essentially this is a shared responsibility between members of the institution:

- It is the policy of UCL that responsibility for managing and preserving research data is shared between all members of the institution.

As a policy position, the document states that:

- It is the policy of UCL that following primary use (e.g. publication) or when research data is archived for long term preservation, these data will be made available in the most open manner appropriate. *Unless covered by third party contractual agreements, legislative obligations[21] or provisions regarding ownership, UCL research data will be provided using a Creative Commons CC0 waiver[22]; supported by data citation guidelines similar to existing publishing conventions. This will ensure that re-used data are unambiguously identifiable and that appropriate credit and attribution is made.*
- The document then identifies roles and responsibilities in UCL, divided as follows:
- Data Creators (students, supervisors and Researchers)
- UCL Research Data and Network Services Executive
- Director of UCL Library Services and UCL Records Manager
- RIISG
- Vice-Provost (Research)
- Provost

For researchers, for example

It is good research practice, and frequently a requirement for grant applications, to plan data management before commencing any research activity. Often this is in the form of a data management plan. It is the responsibility of the individual researcher, or the Principal Investigators if a team of researchers is involved, to generate and execute a data management plan. A template for Data Management Plans can be found on the Digital Curation Centre website.[23]

Researchers should:

- Develop and record appropriate procedures and processes for the collection, storage, use, re-use, access, and retention of the research data associated with their research program;
- Establish and document agreements for research data management when involved in a joint research project, collaborative research or research undertaken in accordance with a contractual agreement;
- Ensure the integrity and security of their data is maintained;
- Be aware of their obligations and potential liability when handling data protected by the UK Data Protection Act (1998);
- Plan for the on-going custodial responsibilities for the research data at the conclusion of the research project or on departure from the university;
- Include Recommendations in Data Management Planning to the Head of Department or research Unit for destruction of research data;
- Include within research grant proposals appropriate consideration of the cost and time implications of data management within grant proposals.

The UCL policy thus sets the framework for positive action. It ensures that arrangements for data curation are aligned with other UCL policies. The Research Data policy itself ensures that a framework has been created for change and achievement in digital curation going forward.

20 See http://tinyurl.com/uclresearchdata
21 See http://www.legislation.gov.uk/ukpga/1998/29/contents
22 http://creativecommons.org/about/cco
23 See http://www.dcc.ac.uk/resources/data-management-plans

Chapter 2: Advocacy

Advocacy to stakeholders in the production and curation of research data

23. There are several barriers to the success of research data sharing:[24]

- Individual contributor barriers
- Availability of a sustainable preservation infrastructure
- Trustworthiness of the data, data usability, pre-archive activities
- Data discovery
- Academic defensiveness
- Finance
- Subject anonymity and personal data confidentiality
- Legislation/regulation

24. The work by the ODE project (Opportunities for Data Exchange) has usefully identified drivers, barriers and enablers in sharing data, from which the above list of issues is taken.[25]

25. To give just one example of the issues listed above from the ODE work:

Driver: Individual contributor incentives

Research data contributors perceive their rewards as:

1) Preserving data for the contributor to access later - sharing with your future self;
2) Peer visibility and increased respect achieved through publications and citation;
3) Increased research funding;
4) When more established in their careers through increased control of organisational resources;
5) The socio-economic impact of their research (e.g. spin-out companies, patent licenses, inspiring legislation);
6) Status, promotion and pay increase with career advancement;
7) Status conferring awards and honours.

Barrier: Individual contributor barriers

Barriers to contributing data may include:

1) Journal articles do not describe available data as a publication;
2) Published data is not recognized by the community as a citable publication;
3) There is a lack of specific funding in grants to address the pre-archive activities for data preservation;
4) There is a lack of mandates to deposit high quality data with appropriate metadata in preservation archives;
5) Journals do not require data to be deposited in a form where it can be re-used as a condition of publication;
6) Data publication and data citation counts are not tracked and used as part of the performance evaluation for career advancement;
7) There is a lack of high status awards to individuals and institutions which contribute data that is re-used.

Enabler: Individual contributor barriers

The barrier to contributing data for publication can be overcome by several proposed solutions:

1) Journal articles describing available data as a publication;
2) Citation of data itself, and the articles describing it;
3) Specific funding in grants to address the pre-archive activities for data preservation;
4) Enforced funding regulation to ensure the depositing of high quality data with appropriate metadata in preservation archives;
5) Journals requiring data to be deposited in a form where it can be re-used as a condition of publication[26]
6) Tracking data publication and data usage and citation counts, and using them as part of the performance evaluation for career advancement;
7) High status awards to individuals and institutions which contribute data that is re-used.

26. More fundamentally, the growth of open data must be underpinned by a shift towards a culture of open access, and the clear articulation of the values and benefits, for the individual, the institution, the research community, and broader society, of open access to data. Advocating the benefits of open data not only starts the ball rolling in terms of a cultural shift, it should also provide funders and decision makers with a strong case for investment in initiatives (such as training and infrastructure) to overcome the barriers to making research data open.

27. Given that open access to research data is in its relative infancy and new practices, infrastructures and policies are constantly emerging, sound advocacy for open data will necessarily be founded in sustained engagement

24 Dallmeier-Tiessen S, Darby R, Gitmans K, Lambert S, Suhonen J, Wilson M (2012). Compilation of Results on Drivers and Barriers and New Opportunities. Retrieved from http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=Compilation+of+Results+on+Drivers+and+Barriers+and+New+Opportunities

25 Ibid., pp. 15-22

26 See the ODE publication listed in n. 24 for a list of references on poor conformance rates

with stakeholders and initiatives on an institutional, disciplinary, national and international level.

Who?

28. In order to foster a cultural shift in favour of open data, advocacy needs to occur at every level within the institution and beyond. All stakeholders can play a role in advocating for open data and all may have legitimate concerns which should be addressed through engagement and advocacy. It is advisable not to take a one size fits all approach to engaging in advocacy for open data. Advocacy means engaging with stakeholders to understand the drivers for sharing data. This will dictate the nature of the advocacy position adopted (e.g. discipline specific, career level).

Leadership

29. Leadership within the institution needs to agree that the institution will wholeheartedly support open data. In order to gain this support leadership and senior management will need to be informed and engaged in the debate around open access to data. There must be buy-in from the heads of faculty for the successful adoption of open data policies. Leadership should ensure that, when developing data management/open data policies, the appropriate stakeholders are engaged i.e. those who will be responsible for its implementation and enforcement, such as the different faculties, library and IT services, the research office and other support departments. It is also up to the leadership of the institution to ensure that data management policies are developed iteratively and that making data open is properly incentivised. This is the route being taken by UCL (University College London), where a Research Data policy has been published. This is being followed by high-level discussions with the Deans of each of the 10 Faculties in UCL on the implications of the policy. In tandem, library and research data staff in the UCL Research Data Service are also being trained in advocacy for the proper management of research data and the benefits of open data. Following this, there will be advocacy meetings with individual researchers and research groups.

Researchers

30. That researchers have concerns about opening up their data has been well documented.[27] These concerns may vary across disciplines, but common issues range from fear of data misuse to loss of competitive advantage. In an environment where researchers are incentivised to value ownership over collaboration and sharing, researchers in certain disciplines have yet to be convinced of the value for them of going to the trouble of making their data open. Following an advocacy model such as UCL's (see above) helps to embed sound research data management practices at an institutional level. There is also potential for researchers to act as advocates for open data; leading by example, through cross-disciplinary engagement to develop standards, and by promoting their own data projects. It is also important for researchers to receive appropriate professional recognition for their engagement.

University support services

31. Libraries are well placed to advocate for best practice in data management and data citation. They can also help to increase the visibility of research data e.g. by acting as a citizen science hub[28] or as a digital laboratory for certain disciplines. Although libraries are experiencing demand for data management support, there is still much advocacy work to be done towards both researchers and infrastructure providers around the role that libraries can play in supporting data management. On the flip side, supporting research data management is a new role for libraries and advocacy is needed to encourage libraries to adapt to this changing landscape.

32. Research Services and Academic IT Services also have an important role to play. IT Services work closely with researchers, and they are often the first place that researchers turn to when they have RDM issues, particularly technological issues.

Infrastructure providers

33. Repositories and disciplinary data archives will need sustainable funding into the future. This will require advocacy efforts at national and international level, directed at policy makers, research funders, research groupings and institutions. Researchers will also need to be encouraged to choose to use this infrastructure and this choice may be based on the perception of trustworthiness of the infrastructure and on its visibility.

27 Gutteridge, C. & Dutton, A (2013) Concerns about opening up data, and responses which have proved effective https://docs.google.com/document/d/1nDtHpnIDTY_G32EMJniXaOGBufjHCCk4VC9WGOf7jK4/edit?pli=1#

28 Lyon, L. (2012). The informatics transform: Re-engineering libraries for the data decade. *International Journal of Digital Curation*, 7(1), 126-138

What?

34. Advocacy efforts can help address drivers and barriers for data sharing. In particular they can address the following:

Rewards of data sharing

35. Data sharing can be incentivised in two ways. The first is by promoting the value and benefits of open data in terms of its value to society and its potential to solve key global challenges, to the reputation of the institution (transparency) and of the researcher (validity of research results). The second is to provide clear incentives to researchers to make their data open. Institutions will have to work on the development of formal policies for promoting and rewarding those generating and sharing data of use to the scientific community and this is a Recommendation of this Roadmap[29]. The promotion of best practice in data citation will help further to ensure that data is cited and hence researchers receive recognition for sharing their data. Case Study 4 in this Roadmap gives examples of how to cite data. Not all data can be open. There may be funding constraints, where use of the data is governed by a pre-existing research agreement. The data may be confidential and as such there may be privacy issues which mean that the data cannot be open. Individual LERU members, or their national jurisdictions, will have policies already in place to govern the non-release of data in such cases. Nonetheless, where data can be made open for sharing and reuse it should be.

Making data visible

36. Good data management, including metadata (data about data), is essential for data to be made visible and reusable. Good data management practices should be promoted as a continuous and integral part of the research process. The perception that data management is technically difficult is best countered with simple guidance on making data open.[30] Data must also be available in the right place and researchers should be informed about, and actively encouraged to use, appropriate infrastructures, such as disciplinary archives. Linking data to publications is another way of increasing the visibility of data and should be encouraged along with the exploration of new methods of publication. Institutional repositories hosting data should make efforts to connect to larger national and international infrastructures to increase the visibility of the datasets they contain and work with international initiatives to adopt best practice for interoperability.

Increasing collaboration and data reuse

37. The main point of making data openly available is so that it may be reused for new purposes. Research projects should be funded solely on the basis of the research they produce. Nonetheless, collaborative projects based on reusing data can be actively encouraged, especially those of a cross-disciplinary nature as these not only uncover new ways of using data, but also foster the development and sharing of best practice across disciplines and potentially facilitate the 'opening up' of cultures within certain disciplines new to open data. Successful data reuse projects can also be held up as examples to illustrate the benefit of and provide a counterpoint to arguments against making data open.

Building trust

38. Researchers need to trust the keepers of their data. Institutional repositories and data centres investing in certification and standards need to translate this activity to the end user to create both trust in the infrastructure and a sense of prestige. Researchers also need to trust that their data will not be misused. This trust can be built by engaging researchers about the specificities of the types of data they are producing and how open data infrastructures and policies can accommodate them.

Recommendations

39. LERU, researchers and research funders should:

40. Engage at international level to build and collect evidence and advocate for the value of open access to research data

41. Foster a debate amongst stakeholders and disciplines around data sharing

42. Develop and clearly articulate incentives for researchers to make their data open

43. Promote best practice in data management, citation and interoperability to increase the visibility of data

44. Institutions should work on the development of formal policies for promoting and rewarding those generating and sharing data of use to the scientific community. This is a new area for the research community, and LERU members could usefully work together to identify best practice in this regard

29 Gorgolewski, K. J., Margulies, D. S., & Milham, M. P. (2013). Making data sharing count: a publication-based solution. *Frontiers in neuroscience*, 7

30 White et al. (2013) Nine simple ways to make it easier to (re)use your data. PeerJPrePrints 1:e7v2 http://dx.doi.org/10.7287/peerj.preprints.7v2

Chapter 3: Selection and Collection, Curation, Description, Citation, Legal Issues

Selection and Collection

45. There is a taxonomy for data, and this is represented in Figure 1 in the ODE Data Publication Pyramid.[31] Data can vary from raw data and datasets, represented at the bottom of the diagram through to data which is contained and explained from within a published article. A full description can be found in chapter 1 of the ODE report, which has attempted an initial taxonomy of data.

46. The main purpose of this pyramid is to explain the different manifestations research data can have in the context of their availability within, with, supplementary to or referenced from an official scholarly article as the main manifestation of the record of science. As yet there is no agreement as to which categories of data should be curated. As such, this decision should be left to the individual researcher or research group. Looking at the Data Pyramid in Figure 1, there will be a strong case for archiving data used in publications or processed data. For data collections and raw data, the researcher should make a decision based on a number of criteria, which are spelled out in Chapter 4 in the section on Data Management Tools. However, further work should be done by the LERU research community to achieve consensus, for all the categories of data outlined in the ODE Data Publication Pyramid, on which type of data can be made available for sharing and re-use, or made open, and which cannot.

47. Data collection should be as much of an automated process as possible, both for reasons of efficiency and in order to prevent any human intervention from falsifying or otherwise manipulating the data. If data cannot be collected in this way, it is at least vital to establish how the data has been processed and why. The rules for data collection should stem from the researcher, but be based (where necessary) on external requirements from e.g. research funders or a journal. The best practice identified in this chapter can usefully support the researcher identify how to select, describe and curate their research data.

48. In terms of efficiency, data collection should ideally be part of an automated process. Such a process could originate at a digital source whenever possible and appropriate, such as CRIS (Current Research Information Systems) environments or digital laboratory journals. In this way automated

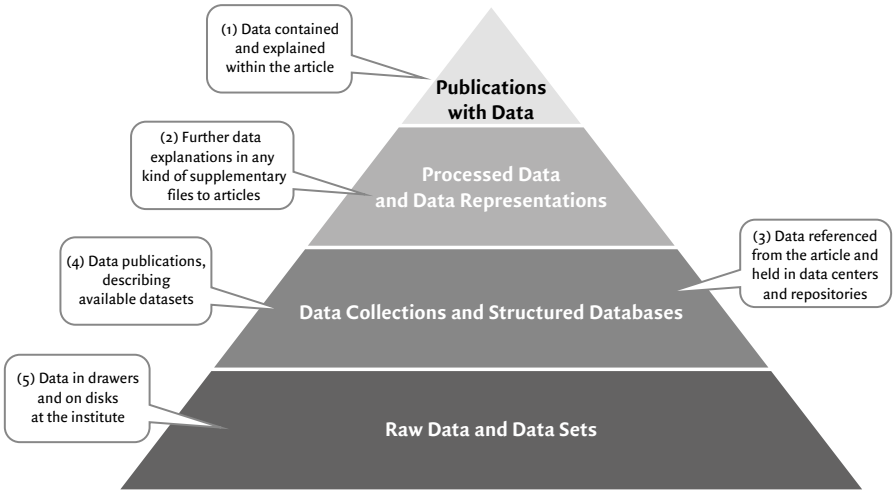

Figure 1: ODE Data Publication Pyramid

31 See http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf

data collection, curation and description could be embedded as an integral part of the workflow. Such developments are new, and LERU institutions could usefully exchange best practice as such developments take root.

49. However, although current research data can mostly be expected to be born digital, some current data as well as most legacy data will come in analogue format and will need to be digitized. Most of these analogue formats contain data in pre-aggregated structures such as tables or graphs – translating these into standardized digital representations that can effectively be processed is non-trivial and still requires research and standardization efforts. Such conversion of data from analogue to digital formats will incur costs, and LERU institutions should undertake a cost benefit analysis to see if such conversion is the best way to spend institutional/ research funder resources.

50. Collection of research data must be conceived as part of a complex workflow (see Figure 2) with the aim of being able to process the data effectively. There are a number of benefits in doing this. It would be possible to check and verify the results stemming from methodological approaches used to analyze the original data. Should a researcher wish to use new ways of analyzing the data, this would also be possible. Such approaches are only possible where the original data is being collected, curated and described.

Data curation

51. For those universities new to data curation, a helpful introduction to the issues can be found in the report From Data Deluge to Data Curation.[32]

52. The PARSE.Insight roadmap for a science data infrastructure focused on long term preservation[33] and conceived the technical components of the infrastructure in terms of what threats they were

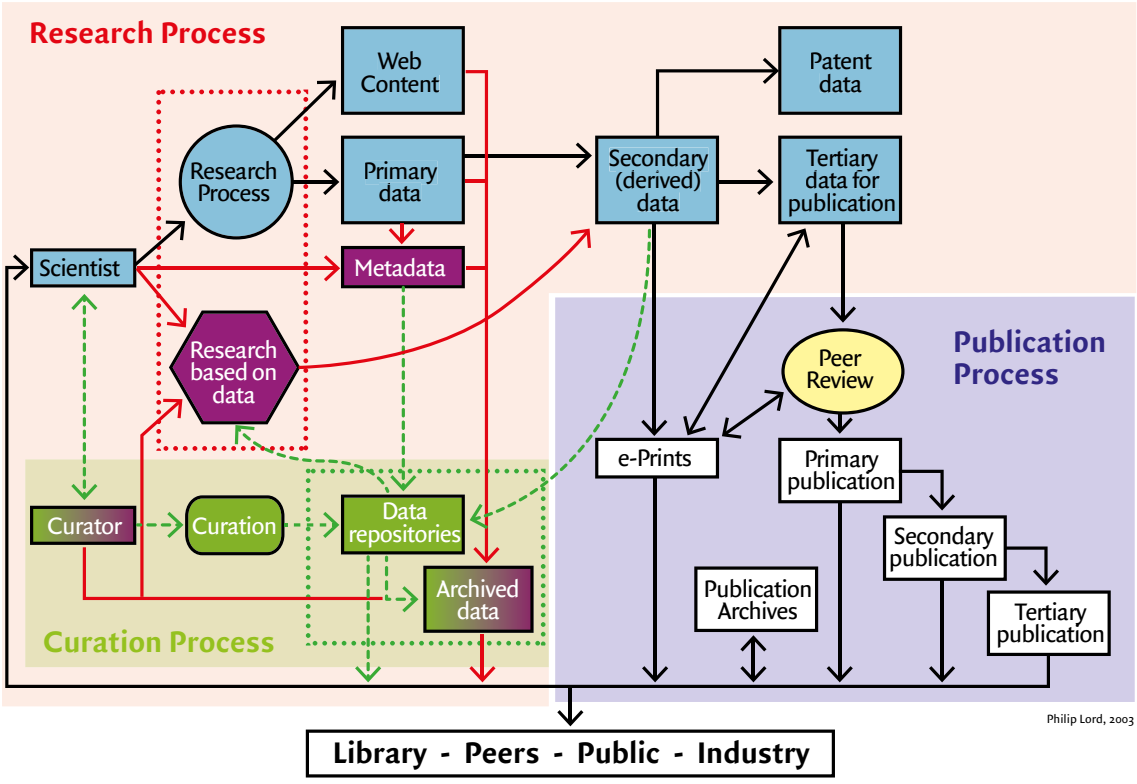

Figure 2: Model of the research workflow

32 P. Lord, A, Macdonald, L. Lyon and D. Giaretta, From Data Deluge to Data Curation, available at http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/150.pdf

33 See http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf

designed to counter: for example, a component to provide evidence of authenticity of a digital object counters a threat that the chain of evidence may be lost and there may be lack of certainty of provenance or authenticity. The PARSE.Insight Roadmap is therefore useful in helping LERU institutions tackle the challenges of data curation.

53. To be useful, open data needs to be reusable in the future. To this end, the data should be properly managed or curated. The curation applies mainly to primary (raw) data or secondary (calibrated and derived) data. Tertiary (analyzed) data are sometimes archived with their publication (see Figure 1). To ensure that data remain accessible, readable, and discoverable, the management process needs to pay careful attention to storage and migration. A mixture of roles and responsibilities is demonstrated below in an analysis by ETH Zurich of their functional requirements for Digital Curation.

What?	Why?	Who?
Data Curation	Ensure intellectual re-usability	Data Producers
Content Preservation	Ensure technical re-usability	ETH Bibliothek
Bitstream Preservation	Ensure technical stability	IT Services, ETH Zurich

Table 1: Digital Curation – Functional Levels[34]

Case Study 2 - Digital Curation at ETH Zurich

Aim of the project
Provision of services for the safeguarding and long term preservation of research data, administrative records and library materials at ETH Zurich.

Description of the project
A recent ETH-wide survey on the handling of research data revealed that many researchers need to be able to record and manage their data in a structured way before it is archived in accordance with the principles of good scientific practice[35] or transferred to a long-term archive. It is important to establish a permanent citation link via a digital object identifier (DOI), in addition to ensuring verifiability, for published data. Data that cannot be recovered is to be made available for long-term analysis or for later use from a different perspective.

Together with suitable partners, ETH-Bibliothek is creating the basis to support researchers of ETH Zurich with the management of their data.

The planned technical solution will also be employed for the long term preservation of administrative records of ETH Zurich and of library materials.

Synergies and context
The current project utilises lessons learned from previous long-term archiving projects for the ETH-Bibliothek, the IT Services department of ETH Zurich, the Consortium of Swiss Academic Libraries and e-lib.ch.

Co-operation partners
- IT Services of ETH Zurich
- Heads of IT Support of ETH Zurich's departments
- Research groups of ETH Zurich
- ETH Zurich University Archives
- A possible co-operation with Zentralbibliothek Zurich is under discussion

Timeframe, sequence

Duration:	2010 to 2013
October 2010:	Start of project
March 2011 to April 2012:	Formulating requirements with the pilot partners
May 2012 to late 2012:	Continued development of the software to be used, in accordance with the requirements
From July 2012:	Practical tests with the pilot partners
From mid-2013:	Gradual introduction of a range of services for all ETH Zurich departments, subject to a financial commitment from the Executive Board

34 See http://e-collection.library.ethz.ch/eserv/eth:4855/eth-4855-01.pdf
35 See https://www.share.ethz.ch/sites/rechtssammlung/Rechtssammlung/4%20Forschung%20und%20wissenschaftliche%20Dienstleistungen/Guidelines%20for%20Research%20Integrity%20and%20Good%20Scientific%20Practice%20at%20the%20ETH%20Zurich.pdf

Description

54. It is good practice in research to ensure that all data generated or collected through the course of research are properly described. Documentation and metadata requirements support this and should be identified from the start of any project and considered throughout the lifecycle of the data. Documentation on research data is relevant at the project level, for individual files and at other levels. Metadata about the research data produced is a core part of a data management plan and of public access.

55. Researchers will commonly describe their data themselves and there is no single standard which defines metadata elements in the description of research data. Researchers are often unprepared for the creation of this metadata. In one study, 42% of researchers surveyed seemed unsure what metadata was appropriate.[36] Librarians have not commonly prepared metadata descriptions for research data. Case Study 3 illustrates how subject communities are beginning to tackle this significant issue. This is an area where librarians and researchers should work closely together to identify best practice and this forms one of the Recommendations of this Roadmap.

56. To ensure that metadata description can be carried out as efficiently as possible, researchers should be able to tap into information and support services where needed, organized by the institution (see also Chapter 6).

57. The administrative and descriptive elements of the metadata will:
- Facilitate interpretation of the data
- Aid resource discovery and re-use (and reduce duplication of effort)
- Provide digital identification (DOI) for citation
- Promote interoperability
- Support archiving and preservation.

58. To proceed in choosing a digital identifier, several issues must be considered: it must facilitate citing and referencing the data on the Web, and participate in the process of making them durable. DOI is one identifier among others, some may be more suitable. As promoted by the Linked Data method, URIs (Uniform Resource Identifiers) can be used as unique and unambiguous data identifiers, enabling the sharing and re-use of research data.

59. Metadata should be created efficiently (only once) and re-used for different purposes. Therefore,

Case Study 3 - Metadata management in subject repositories

From Research Data Management: Practical Strategies for Information Professionals, ed. J.M. Ray, (West Lafayette: Purdue University Press, 2014), in press, p. 151. Disciplinary repositories do often offer help with metadata for deposit, however. The Inter-university Consortium for Political and Social Science Research (ICPSR)'s data preparation guide for its social science data repository includes a section on "best practice in creating metadata," focused heavily on using the Data Documentation Initiative (DDI),[37] within its Data Preparation Guide.[38]

The Dryad repository for the basic and applied biosciences[39] provides a relatively short set of submission guidelines for researchers, including Recommendations clearly to indicate column headings, document symbols indicating missing data, use of the ISO8601 date format, and use of taxonomic or other standard names when appropriate. Dryad documentation also goes a step further to provide a two-minute YouTube video demonstrating the submission process.

The Odum Institute's node of the Dataverse repository[40] for social science data takes a different tactic, simply stating that the preference is for "fully documented" data from data analysis packages with all supporting interpretive information. The Odum Institute provides some tools for collecting information, requiring a five-page deposit form that provides some information about the structure of the data contributed. It also offers assistance to researchers in preparing data that may not yet have adequate documentation.

36 See Research Data Management: Practical Strategies for Information Professionals, ed. J.M. Ray, (West Lafayette: Purdue University Press, 2014), in press

37 See http://www.ddialliance.org/

38 See http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/index.html

39 See http://datadryad.org/

40 See http://arc.irss.unc.edu/dvn/

interoperability between existing CRIS-systems, data curation systems and discovery tools (by adherence to standards) is crucial. It is recommended that researchers receive support to describe their data from the librarians and IT support staff.

Citation

60. Datasets are increasingly a significant part of the scholarly record and are being published more and more frequently, either as part of a publication, or on its own. Data citation acknowledges the author's sources, makes it easier to find data, promotes the reproduction of research results and makes it possible to recognize and reward the data creators. The basis for all this is standard citing methods.[41]

61. Standards for the citation of data are not uniformly agreed upon yet. However, many data providers and distributors and some style manuals do provide guidelines, and several initiatives have been taken to develop bibliographic standards for data[43].

62. Typically, in addition to responsible party, title, name of repository, analysis of software, data accessed, effective citation would include a digital (persistent) object identifier or DOI[44], and a link to help users get to the data directly.

63. It is recommended that institutions inform researchers about what is expected of them in this area. Publication of 'basic principles', general guidelines and examples to help researchers how to cite data are very common and helpful. This information should include links to existing demands from funding agencies, publishers and data centers, specific to the different disciplines. LERU could encourage more collaboration and sharing of effort in this area.

Legal issues

64. The purpose of this section is to explore licensing arrangements which are suitable for research data, databases and datasets.

65. In general, research data is an object that cannot be protected individually by copyright under the current European legal framework.[45] However when data is set together creating a database, then this aggregation can be protected in two different ways: if the selection, arrangement or presentation of the elements is original, then a database can be pro-

Case Study 4 - how to cite Research Data

From DataCite recommendations:
Creator (Publication Year): Title. Publisher. Identifier (see http://www.datacite.org/whycitedata)

Exemplified by:
Piguet, Bruno; Legain, Dominique; (2011): Tethered balloons CNRM Site 1; Météo-France, GAME. http://dx.doi.org/10.6096/BLLAST.TETHEREDBALLOONSCNRM

This example was used in the ODE report on best practice for the citability of data.[42]

This is an example of citing different layers of data from PANGEA:
Citing data from a specific cruise:
Haardt, H; Maaßen, R (1983): Physical oceanography from the Drake Passage and Bransfield Strait during Meteor cruise M56. Institut für Angewandte Physik, Christian-Albrechts-Universität, Kiel, http://dx.doi.org/10.1594/PANGAEA.737666

Citing a single data profile from a cruise:
Haardt, H; Maaßen, R (1983): Oceanographic and optical profile at station M56_127-235. http://dx.doi.org/10.1594/PANGAEA.80634

41 For a Draft Declaration of Data Citation Principles, by the Research Data Alliance, see http://www.force11.org/datacitation

42 See http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=Report+on+Best+Practices+for+Citability+of+Data+and+on+Evolving+Roles+in+Scholarly+Communication

43 See for an overview: Green, T (2009), "We Need Publishing Standards for Datasets and Data Tables", OECD Publishing White Paper, OECD Publishing. doi: 10.1787/603233448430 http://dx.doi.org/10.1787/603233448430

44 In 2012, DOI became part of ISO Standard 26324

45 For the European Copyright Directive, see http://en.wikipedia.org/wiki/Copyright_Directive

tected like any other creative work by copyright. Although there is a lack of originality, databases can also be protected by a sui generis right that rewards the creator of the database for the investment of time, effort, energy or even financial resources.[46] The term of protection is different in each case: as an original work a database is protected all the life of their creators plus 70 years and under sui generis rights it is protected for 15 years after its creation or its publication. Therefore, it is important to determine who the rights holder of the database is - researchers, universities, or funders – if data is to be shared and reused.

66. To express the terms of reuse, it is recommended to use a suitable open content licence that includes a mention of databases rights[47] because in some licences there is no mention of sui generis rights. It is advisable that researchers and institutions are aware of these specific licences and the requirements by funders to use them.

67. In other cases, for instance images, research data can be considered as any other original work. Therefore it is not enough to establish an open licence in a database. In those cases, it is necessary to establish a similar licence in the data itself by using one of the existing licences.

68. Beside copyright, creators of research data may be under ethical, legal and/or contractual obligations to protect their data, ensuring respondent and researcher anonymity and respondent consent. Issues around confidentiality and intellectual property should be considered when (more or less restrictively) sharing data. When database rights belong to universities, they should include openness by default including opt out situations when sharing of data can involve privacy or confidential issues. Individual LERU members, or their national jurisdictions, will have policies already in place to govern the non-release of data in such cases.

69. Institutions can include a general statement on legal issues in their Data Management policy. However, because of the high level of complexity of the issues involved, practical support for researchers is even more crucial. Legal experts, for example the university's legal office, should be prepared to give advice on relevant contracts, regulations and legislation, in addition to information on the website and presented in training sessions. Moreover, many university libraries have created teams to support researchers that can help with these openness issues due to their expertise in open access. There may be resource issues here in terms of new staffing posts, or skills and knowledge needed by existing staff. Collaboration between LERU institutions would help to address these issues.

70. Although legal issues around research data are global in scope, laws about data and databases may vary across countries. It is important to know the specific legal framework in each case. These are all part of responsible data management practice.

71. The potential of Text and Data Mining (TDM) is acknowledged by researchers, who see the benefits of using automated tools to mine the literature and supporting research data.[48] The use of Text and Data Mining tools underpins the Royal Society Report Science as an Open Enterprise.[49] However, the legal basis to allow the use of TDM techniques, certainly in licensed commercial literature, is unclear. What is needed at a European level is a Fair Dealing Exception, certainly for the purposes of research, in the EU Copyright and Database Directives to facilitate the sharing and re-use of research data.

Recommendations

72. LERU institutions, researchers, research funders and the bodies of the EU should consider the following:

73. Further work should be done by the LERU research community to achieve consensus, for all the categories of data outlined in the ODE Data Publication Pyramid, on which type of data can be made available for sharing and re-use, or made open, and which cannot.

46 For the European Database Directive, see http://en.wikipedia.org/wiki/Database_Directive

47 For instance, the Open Knowledge Foundation offers three different licences http://opendatacommons.org/licenses/ while Creative Commons recommends CC0 for databases http://creativecommons.org/about/cc0

48 For an introduction to Text and Data Mining, see the LIBER Factsheet at http://www.libereurope.eu/sites/default/files/Text%20and%20Data%20Mining%20Factsheet.pdf

49 Available at http://royalsociety.org/policy/projects/science-public-enterprise/report/

74. Institutional policies and practices for data curation need to be rooted in the Best Practice identified in this Roadmap.

75. Documentation and metadata requirements should be identified from the start of any project and considered throughout the lifecycle of the data.

76. Metadata should comply with existing standards for the content. Chosen formats should preferably support machine-to-machine interoperability. Interoperability between existing CRIS-systems, data curation systems and discovery tools is crucial. Adhering to standards will provide a basis for this.

77. Institutions, librarians and researchers should work together to clarify what is expected of researchers when describing and citing data. Publication of 'basic principles', general guidelines and examples to help researchers how to describe and cite data are very common and helpful. This information should include links to existing demands from funding agencies, publishers and data centers, specific to different disciplines.

78. It is important to identify the owner of the data: the researcher, funder or institution. Responsibilities for stewardship of the data both during a project (if the work is project-based) and when funding has come to an end should also be clear. In cases of multi-party research projects (for example 7 university, 2 business and 3 government agencies working on one project) the partnership agreement which underpins the collaboration before the research starts should identify how resulting research data will be managed and who owns it.

79. To express the terms of re-use for datasets, it is advisable to use an open content licence suitable for data. Researchers and institutions should be made aware of such legal tools.

80. Practical support for researchers should be organized. Legal experts, for example the university's legal office, should be prepared to give advice on relevant contracts, regulations and legislation, in addition to information on the website and presented in training sessions. Librarians with expertise in offering digital information services could offer their expertise in open access too to support researchers.

81. To facilitate Text and Data Mining, European copyright frameworks in the EU Copyright Directive and the EU Database Directive need to be revised.

82. Institutions need to be aware that there are resource implications in developing a responsible approach to research data management. Costing the true costs of research data management is in its infancy. Chapter 5 gives a summary of the current state of knowledge and practice in costing such developments.

Chapter 4: Research Data Infrastructure

Context

83. The growth of research data presents an infrastructural challenge for researchers and for universities, a decade ago termed the 'data deluge'.[50] The infrastructure required to deal with the increasing amount and importance of research data comprises four aspects:

i. Research data – Present in different formats and depending on specific research methodologies.
ii. Data management tools – Needed for creation and analysis, administration, documentation, archiving, publication and discovery.
iii. Technical components – Based on local desktops, servers or cloud services.
iv. Staff – Involving, e.g. managers, system administrators, curators, support staff.

Research Data

84. Research data cover a wide range of formats, from spreadsheets or databases for social science surveys over spectroscopic images or genomic data to digitized images of ancient books. Display and analysis of research data often depends on specific computer programs, thereby requiring not only the data object itself but further facilities such as specific software, servers and documentation of workflows that describe the use of these resources. These facilities will differ with the research methods applied, for example:

a. **Quantitative**, often based on statistics and requiring bespoke online software environments and processing power, sometimes 'supercomputing'.
b. **Qualitative**, such as video-interviews often involving massive storage requirements.
c. **Hermeneutic**, for example contextualizing cultural artifacts involving typically textual and image resources, e.g. books or paintings, interlinked in semantic networks that require elaborated solutions for metadata management and advanced discovery technology.

Recommendation

85. Research data infrastructure needs to offer a generic framework for a wide variety of research processes and outputs to create, process and share data, e.g. quantitative, qualitative and hermeneutic methodologies.

Data Management Tools

86. Research data infrastructure typically involves many tools, an individual description of which is beyond the scope of this paper. An overview can be provided in terms of functional areas.

a. **Creation and Analysis**: Instruments for research data capture, the software that supports them as well as storage and analytics.
b. **Administration**: Data management planning in the project preparation phase and the fulfilment of reporting obligations as well as funders' policies throughout and after the project. These tools require campus integration with systems for research support, finance and HR (Human Resources).
c. **Documentation**: Descriptive metadata can partly be generated automatically during the data creation phase. But the specifications of workflows and methods, metadata on creators, the structure and semantics of the data have to be added separately. These data are essential to allow later discovery, retrieval and re-use of data.
d. **Storage, archiving and publication**: Not all data can or should be kept for the long term. A selection and appraisal process for data to be retained has to take place, involving the definition of curatorial terms and conditions. If applicable, the publication, including the assignment of persistent Internet addresses (e.g. PURL, DOI), has to be arranged. In some cases, anonymization has to take place or data have to be technically protected for security, privacy or legal reasons. If research is externally funded, a funder may well set down re-

50 Hey, A. J. G., & Trefethen, A. (2003). The Data Deluge: An e-Science Perspective. In Grid computing: making the global infrastructure a reality (pp. 809–824). New York: J. Wiley. Retrieved from http://eprints.soton.ac.uk/257648/

quirements for the length of time data needs to be retained after funding has come to an end. If there is emerging good practice in a particular subject discipline on rules for data rentention and disposal, then these should be followed. In many cases, however, good practice will not yet be available and the researcher/research group will have to make a decision and include that decision in their data management plan. Looking at the Data Pyramid in Figure 1, there will be a strong case for archiving data used in publications or processed data. For data collections and raw data, the researcher should make a decision based on a number of factors, including costs in archiving data, the importance of making this type of research data available to the wider research community, the availability of suitable data curation infrastructure and services to undertake this activity. Further work should be done by the LERU research community to achieve consensus, for all the categories of data outlined in the ODE Data Publication Pyramid, on which type of data can be made available for sharing and re-use, or made open, and which cannot.

e. **Discovery**: Re-use and acknowledgement, e.g. citation, can only be achieved if research data can be found in search engines such as Google (Scholar), subject-specific discovery services or institutional data catalogues. Universities need to make sure that they keep a record of those research data archived or published – at least those underpinning research publications.

87. Data management is an established part of the research process and parts of the infrastructure have already been developed, particularly as subject specific services. For example, Earth Sciences have built a system of World Data Centres, Life Sciences rely on services such as those from the National Center for Biotechnology Information (NCBI) or the European Bioinformatics Institute (EMBL-EBI) and High Energy Physics have large facilities such as CERN. Universities have long provided storage and computing facilities. Data curation and long-term preservation services are increasingly provided in academic libraries. A model of good practice in the biosciences can be found in the Case Study Making the Best Use of Life Science Data, which is included in one of the publications from the ODE project (Opportunities for Data Exchange).[51]

88. The growing importance of research data in all research fields – especially those fields not having their own solutions – increases the demand for institutional infrastructure. Indeed, the 'long tail' of research data[52] residing on local desktops, hard disks and servers might well comprise a bigger challenge than 'big data' where facilities are externally available. Additionally, funders increasingly introduce policies requiring researchers and institutions to keep and open up research data produced with their funding. Thus, a 'long-tail' data repository and a tool facilitating management and reporting, e.g. a data catalogue, are primary candidates for institutional responsibilities.

Recommendation

89. A portfolio of tools for an institutional research data infrastructure that fills the gaps of existing external research infrastructures should be developed at LERU universities, e.g. a 'long tail' data repository and a data catalogue supporting re-use and open data. To save costs the sharing of services should be considered.

51 See http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/10/7782_ODE_Brochure_v5.pdf, pp. 6-7

52 P. Bryan Heidorn. (2008). Shedding Light on the Dark Data in the Long Tail of Science. Library Trends, 57(2), 280–299. http://dx.doi.org/10.1353/lib.0.0036

Case Study 5 – Dataverse Netherlands

Based on R.T.A. Grim, Tilburg University

Aims

- A shared research data infrastructure for Utrecht University, Tilburg University, Erasmus University Rotterdam, Maastricht University, University of Groningen, 3TU Datacentrum and the Netherlands Institute of Ecology (NIOO)
- To make the datasets underlying the scientific publications available, accessible and findable for reuse.
- To support researchers with research data management during all research phases.

Who finances it, what is the duration?

- Usually, central funding is available within each institution for research data storage in Dataverse. When the central resources are depleted, researchers contribute to the costs for data storage. Contracts are renewed on an annual basis. Partner institutions might use different internal cost models.

Who is responsible/co-ordinates?

- Utrecht University is formally responsible for the co-ordination of Dataverse Netherlands, governed by a working group in which all partners participate.

What kind of research data is targeted?

- The main focus is on the research data sets and supplementary materials which lie behind scientific publications
- Raw data might be archived in a "dark Dataverse". A "dark Dataverse" is an archive that is not released and is therefore not findable or accessible, unless you have access to the URL and have access permissions. A dark archive might be used to archive datasets that cannot be released due to privacy regulations or any other legal directives.

How is local support organized?

- Each university and faculty has its own Dataverse. Within each Dataverse, collections are defined for departments and research groups in some cases, in other institutions each researcher or research group uses a separate Dataverse.
- Partner institutions can use a different support model for research data archiving.
  - For example, institutions may allow researchers to create a Dataverse (collection) themselves. Other institutions make use of a predefined structure for a Dataverse collection. At Tilburg University a Dataverse collection is always defined at "group level", i.e. a research department or a research group.
  - Institutions also differ with respect to how DVN is used in the research lifecycle: Tilburg University for example by preference archives the final version of the dataset that was used for the paper. Utrecht University allows researchers to use Dataverse during all phases of research.
- To co-ordinate access to and the functional management of the Dataverse Netherlands three types of documents are needed: 1. general terms of use, 2. service level agreement (SLA) and, 3. a document which details the functional management of the DVN application.

What quality measures are in place?

- The requirements set by the Data Seal of Approval are used as a general framework for data quality and procedures. At this point in time, each institution is free to decide whether or not to apply the framework.
- Each research group assumes ownership for research data quality. Data quality requirements and conditions for data audits, checks and balances are described in specific data management plans (DMPs).

What restrictions on access are in place? What licences are used?

- Specific access conditions for each publication might apply to the datasets and supplementary materials. "General terms of use" that can be used by the institutions that participate in Dataverse Netherlands are being prepared. In addition specific terms of use might be applicable.

What kind of support services are provided to researchers?

- Library staff actively support the following activities: archiving of research data and supplementary materials, the registration of datasets, support for data curation.

Website

- The Dataverse installation in The Netherlands was originally started in Utrecht University Library for Utrecht University and then adopted by other institutions.
- The above Case Study was originally elaborated by LIBER's E-Science Working Group.
- Dutch Dataverse Network (v3.3): http://www.dataverse.nl/dvn/

Technical Components

90. Many institutions have carried out preliminary "state-of-the-nation" surveys, questionnaires or audits to assess the range, volume and status of existing datasets within departments and faculties.[53] Tools such as the DCC Data Audit Framework or the DCC CARDIO tool or Bristol Online Surveys have proved useful in this context. The results have been synthesised into data curation requirements reports, which have then informed service planning.

91. Technical components of research data infrastructure are distributed across the university:

- **Local** desktops as well as departmentally-located servers and hard disks play a crucial role in supporting specific research methodologies at the individual or group level, e.g. by linking data capture devices to storage facilities and deploying bespoke data analysis software.
- **Central** infrastructure in the form of Virtual Machines and storage, sometimes supplemented by high performace computing facilities, allow an efficient management of Internet-based data management applications and resilient multi-site deployment as well as supporting the durability of research data through back-up, replication and tape archiving.
- **Shared** infrastructure between universities or research organisations as in the case of Grid computing or joint research data networks is based on location-independent, network-based deployment. Dedicated 'Cyberinfrastructures' (US terminology) or 'e-Infrastructures' (EU terminology) are often subject-specific or related to data generating facilities such as the worldwide network managing the data for the Large Hadron Collider at CERN.
- **Outsourced** operations are often based on cloud computing and allow universities to provide computing, storage and archiving without their own hardware and virtualized infrastructure by 'renting' it from commercial providers such as Amazon (e.g. EC2, Glacier). However, this might be more expensive than local deployment of large-sized operations.[54]

92. The right distribution of technical components and their operational control across the university is a trade-off between the availablity of local expert support and the prevention of costly redundancies. This task is further complicated by the continuously-changing requirements for research data infrastructure.

Recommendation

93. LERU universities should establish an asset register of local, central, shared and outsourced research data facilities, allowing continuous monitoring of the efficiency of operations and a strategic approach to capital planning for the 'digital estate'.

Staff

94. The people involved in the management of research data, the tooling and the techical components have varying profiles. As with tools, the staff are distributed, in local departments, central services and outside the university:

- Researchers are the individuals fully understanding the requirements for infrastructure in their specific research field. Departments often deploy local IT Officers.
- System administrators in central IT services bring the skills for managing generic storage and internet operations.
- Data curators or subject specialists in libraries look after consistency and persistence as well as terms and conditions of valuable research data, when the researcher has finished a project or has left the university.
- Research facilitators in administration provide the required knowledge of policies and processes.
- Academic IT advisors can support researchers in their choice of appropriate technologies and standards.

Recommendation

95. To minimize the time researchers have to spend on technical and administrative processes, LERU universities should organize the institutional research data workforce of local IT Officers, subject librarians, central system administrators and research facilitators into a coherent support service, bringing them together to provide a full service for researchers with co-ordinated provision for information, guidance and training.

53 For an example from the University of Leeds in 2012, see http://library.leeds.ac.uk/blog/roadmap/post/152

54 Rosenthal, D. S. H., & Vargas, D. L. (2013). Distributed Digital Preservation in the Cloud. International Journal of Digital Curation, 8(1). http://dx.doi.org/10.2218/ijdc.v8i1.248

Conclusions and Recommendations

96. Research data infrastructure is an essential prerequisite for today's and tomorrow's research. While several services exist within the research communities and outside the university, the increased amount and importance research data reinforces demands for university facilities – also because funders expect researchers to make their data openly available in the long-term. Research data infrastructure is thus to be seen as an institutional asset in the digital estate.

Recommendations

97. LERU institutions should note the following Recommendations:

98. Research data infrastructure needs to offer a generic framework for a wide variety of research processes and outputs to create, process and share data, e.g. quantitative, qualitative and hermeneutic methodologies.

99. A portfolio of tools for an institutional research data infrastructure that fills the gaps of existing external research infrastructures should be developed at LERU universities, e.g. a 'long tail' data repository and a data catalogue supporting re-use and open data. To save costs the sharing of services should be considered.

100. LERU universitites should establish an asset register of local, central, shared and outsourced research data facilities, allowing continuous monitoring of the efficiency of operations and a strategic approach to capital planning for the 'digital estate'.

101. To minimize the time researchers have to spend on technical and administrative processes, LERU universities should organize the institutional research data workforce of local IT Officers, subject librarians, central system administrators and research facilitators into a coherent support service.

Chapter 5: Costs

Context

102. The revolution that data-driven science has initiated presents great challenges for a university and its finances, particularly at a time when many European countries face very significant fiscal challenges. University budgets are under considerable pressure and this makes the identification of costs in supporting research data management of significant importance for university planning. Alternative funding sources could be the EU and individual research funders, although not all costs (such as recurrent staffing costs) would be considered as eligible costs by external funders. The extent to which research funders will fund the storage of, and access to, research data after the end of a project is also a factor to be taken into account when costing the construction and sustainability of research data infrastructures.

Cost Models

103. There is no one single model which is generally used to calculate costs. One of the most influential stems from two studies funded by the JISC (Joint Information Systems Committee) in the UK. This model comprises three main elements:

1. Key Variables and units which affect costs
2. Activity model for research data, identifying activities with cost implications
3. Resources which have a bearing on cost

104. Full information on what activities are contained in each of these elements can be found in the two KRDS (Keeping Research Data Safe) studies.[55] The KRDS approach to costing is one which could usefully be adopted by LERU members

105. All these costing activities are linked to TRAC (Transparent Approach to Costing) in the UK, which allows the Full Economic Costs to be identifed.

55 N. Beagrie, J. Chruszcz, B. Lavoie and M. Wollard, *Keeping Research Data Safe* (2008 and 2010); see http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx and http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx (last accessed 10.8.13). JISC is the body overseeing the compilation of the Reports; KRDS is the acronym by which these two costing studies are known; and TRAC is a methodology for calculating the full economic costs of research

Case Studies

106. Given that research data management is a new discipline, the best way to gain a handle on costs is to look at case studies in LERU universities which are undertaking activity in this area.

University of Oxford

107. In May 2013, proposals were considered in Oxford for a suite of Data Management Services and projects to help establish them. Over 2 years, the costs were estimated at £2,118,000. These costs covered DataFinder (to describe and catalogue datasets for retrieval), DataBank (for storage and preservation of data), programme co-ordination, Storage-as-a-Service, and ORDS (online research database service to manage and publish relational databases online). Over half the costs were predicted to be incurred in the first two elements of the suite of RDM services. Recurrent costs for the whole suite were predicted as £0.5 million a year,[56] although this number is contingent on the level of uptake (and storage demands) of DataBank.

Case Study 6 - UCL financial case study for Research Data

For UCL (University College London), it was the EPSRC policy on research data which underlined the need for UK research universities to take action on research data. Two of the principles are of particular importance: firstly, that publicly-funded research data should generally be made as widely and freely available as possible in a timely and responsible manner; and, secondly, that the research process should not be damaged by the inappropriate release of such data. A similar approach is being adopted by other research funders.

In terms of costs, EPSRC had this to say. EPSRC recognises that systems and infrastructure appropriate to the storage and management of access to research data have associated costs. EPSRC believes that where research has been publicly-funded it is reasonable and appropriate to use public funds also to fund the associated data management costs. EPSRC therefore expects research organisations to make appropriate provision from within public research funding received, making use of both direct and indirect funding streams as appropriate.

It may be that in the interests of efficiency a research organisation wishes to appoint a third party to provide appropriate services, or two or more research organisations may wish to collaborate and develop a shared service: such approaches would be entirely acceptable within this framework.

UCL took the view that the best way to address this challenge was to create an institutional UCL Research Data Service. UCL's model parameters split the costs into several parts.

- Near term data storage: being commissioned and used now (costs an average of £400/TB total cost of ownership, TCO, good for five years at least)
- Data Curation: No real cost for this yet (but see below)
- Digital preservation: This is split as:
  - Bitstream preservation: essentially automatic monitoring for data and media corruption (outsourced to third party and likely in the region of £150/TB/year and front loaded for a minimum of 10 years)
  - Data preservation: more costly and time consuming format monitoring, standards mapping and migration activity (there are no estimates for this yet)

For indicative capital setup costs, the UCL Research Data Service required in the region of £1 million to be established. Going forward the recurrent costs have been estimated to be £500,000 a year, which includes recurrent staff costs.

The scope of the service is that it will engage with funded project research in UCL and provide digital storage for their research outputs. There will be partial cost recovery, with users being charged fractional units of TCO cost, although as storage technology is rapidly changing the unit cost will ultimately depend on cost at time of purchase. As an example, the current average unit cost of £400 per terabyte would most likely be charged as a fractional yearly allocation of £80 per terabyte per annum. Initially, this has the potential to generate income of around £150,000 per year. Phase 2 of the project, which is yet to be implemented, will move from digital storage to longer term data preservation. It is likely that the tech-

56 See http://damaro.oucs.ox.ac.uk/docs/Damaro%20RIMSC%20report%20May%202013.pdf, especially slide 21 (last accessed 18 August 2013)

nology layer will be outsourced to a third party provider, with a service layer being provided from within UCL. This development has not yet been costed in detail, but indicative quotes have suggested that £150 per terabyte per annum for 10 years is a realistic figure and period of preservation (equating to an upfront cost of £1500 per terabyte).

UCL Library Services will take on a role in helping to curate the outputs of Small Science, typically where the researcher is not funded by an external funder. The outputs will reside in the Library's Digital Collections service. This service already exists and will expand as demand grows. Current running costs are absorbed in the Library's recurrent operational budget, but this will change as the service develops. Currently there is no charge to users to deposit their data outputs in the service.

A UCL Research Data policy[57] was approved in the Summer of 2013. This policy spells out for all stakeholders what their obligations are, including financial obligations. For researchers, for example, it stresses that they

- **Include within research grant proposals appropriate consideration of the cost and time implications of data management within grant proposals.**

The UCL Research Data policy therefore recognises that secure research data management, including its financial aspects, is a collaboration between central administrative Departments/Divisions in UCL and individual researchers and research groups. Working in partnership, a sound financial model can be created to sustain the ongoing services.

108. Nonetheless, costing is still in its infancy. The 2010 JISC Report noted the findings at a JISC Workshop attended by a wide range of different types of universities to discuss storage services. The spectrum of costs was between £450 per TB for a single copy for 1 year through to £5000 per TB fully paid-up for indefinite storage using multiple geographically-dispersed copies of the data.[58]

Cost Benefits

109. There are cost benefits to curating research data, which can sometimes help provide a framework for judging the cost effectiveness of research data curation in the first place. The following example from the UK Data Archive illustrates this point well. The UK Data Archive[59] consists largely of unique data, which cannot be replicated:

> Consider the annual General Household Survey. The 2001 wave of this survey told us, amongst other things, that household size was declining slowly, that the prevalence of home ownership and cigarette smoking was flattening out, male employees were less likely to have an employer's scheme pension, but female participation in the same schemes was increasing (Walker et al 2002).The cost of the creation of this dataset is subsumed within a total cost of the GHS (in 2001) which was reported by the National Statistician as being £1.43 million. This figure covered "analysis and reporting for 2000-01, fieldwork for 2001-02 and planning and preparation for 2002-03." (UK Parliament 2001). We can reasonably estimate that the replacement cost for this dataset would be over £500K, but since the results of any replacement would be relating to a different period in time, it would only be a replacement rather than a recreation.

110. Given the uniqueness of the data, the high cost of recreating it, and the impossibility of replicating the original dataset, the cost benefits of curating the orginal dataset can be said to be high.

Small Data

111. In some areas of research, the data outputs are small in relation to those from some areas of science, such as High Energy Physics. In UCL (University College London), the UCL Research Data policy gives UCL Library Services a role through its Digital Collections service. Such a service is in its infancy, and currently there is no extra cost to researchers to deposit their data, although this may change as the

57 See http://tinyurl.com/uclresearchdata (last accessed 18 August 2013)
58 See http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx, p. 34
59 For this and what follows, see http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf, p. 72-4

service grows. In The Netherlands, the Dutch Dataverse Service already provides storage and sharing facilities for researchers (mainly) in the Humanities and Social Sciences at Dutch universities and research institutes.

112. There are also commercial services who make a fee-based offering. Dryad is concerned with data linked to publication and they have a rich price offering.[60] For individuals, the cost would be $80 per data package. As a subscription, for members, the annual fee would be $25 per published article.

Who pays?

113. Who is responsible for paying for the necessary infrastructure for research data management? Many research funders mandate research data management and research data management plans as a condition of funding. Research Councils UK, for example, have published a set of 7 Principles which underpin research data management, one of which says:

> It is appropriate to use public funds to support the management and sharing of publicly-funded research data. To maximise the research benefit which can be gained from limited budgets, the mechanisms for these activities should be both efficient and cost-effective in the use of public funds.[61]

114. Some costs will be eligible as direct research costs on a grant, for example creating and curating a dataset as a project output, so that it can either be curated in-house or the curation outsourced to an external provider. Much depends of course on a university's budget model.

115. Some activities, such as long-term curation, will use infrastructure which is shared in the institution, so that it is not specific to an individual project. Usually, these costs are included in an institution's overheads and should be recoverable through indirect costs on a grant.

116. Long-term curation is sometimes provided by national entities, e.g. DANS and 3TUDatacenter in The Netherlands. In this case, national funding is available (which will cover part of the costs).

117. If external funding is not available, this leaves the question of the original capital investment needed to construct the infrastructure in the first place. It is likely that these costs will have to be met by the institution itself, usually from its research budget. However, the evidence of the JISC studies shows that the costs of archiving activities (archival storage, preservation planning and related actions) are consistently a small proportion of a research institution's overall budget.

Recommendations

118. Cost modelling for research data management is still in its infancy, although good work has been undertaken to date. LERU institutions should continue to add to the growing body of best practice in this area, which will help identify how research data infrastructures can be made sustainable.

119. LERU universities could exchange information on costs using these tools to build up a knowledge-base to inform their development.

120. LERU universities, who decided to band together to construct shared data management services, would be innovative in modelling the costings for such a shared service.

121. The CIO community in LERU has an important role to play in collecting and sharing information on costing for research data management and could act as a focus for best practice in this area.

60 See http://datadryad.org/pages/pricing

61 See http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx (last accessed 10.8.13)

Chapter 6: Roles, Responsibilities and Skills

Introduction

122. It is acknowledged that roles and responsibilities need to be clearly delineated for successful data management and sharing.[62] From leadership and senior management, to support services such as IT and library services, to, of course, the researcher, all are stakeholders in the open research data environment and all play a role in ensuring that open data is integrated into the way research is carried out in the future.

123. In order for responsibilities to be fulfilled stakeholders must be equipped with the skills and education necessary to meet the requirements of their roles. If we are truly to realize the value of making data open, then investment will need to be made in the development of data scientists who are capable of exploiting and reusing data from across domains.

124. Good communication across roles will be key, as will engagement with external networks to ensure interoperability, encourage cross-disciplinary work, and increase the visibility of the data being produced within the institution.

Roles & Responsibilities

125. The first step for research institutions is clearly to outline roles and responsibilities for making data open in their data policy. These could look something like Table 2 below.

126. The researcher is at the heart of the data lifecycle and it is the researcher's responsibility to ensure that data is made open and that good data management practices are implemented from the outset of a research project, e.g. through the creation of a data management plan that is in line with the requirements of the policies of the research funder. At the other end of the data lifecycle, researchers will also be the ones who reuse data and have a responsibility to use appropriate data citation to ensure that the creator of the data receives recognition.

127. It is the responsibility of the institution to ensure that the researcher has access to the support and infrastructure necessary to make data available, whether that be external, e.g. through disciplinary data centres or internal, e.g. through institutional repositories. This support not only includes the infrastruc-

Roles & Responsibilities	
Researcher / Data owner	Data producer/owner (principal investigator) & end user. Responsible for data management plans, for making data open (depositing, data management, choice of software), reuse, retention and relegation
Data Scientist / Data Steward	Works in close collaboration with scientists to collect, exploit and analyse, reuse data, part of the research team. Technology watch. Has responsibility for making decisions about the data, most importantly 'post project', e.g. access, queries, retention
Library	Support for data management & discovery: curation, preservation, data publishing and archiving and access to data resources. Guidance on finding and assessing data, IP, open access licensing, data citation, data management plans. Technology watch
Management, faculty, administration	Policy development and communication, awareness raising, enforcement, education, cutural change
External service provider (data centres, cloud services)	Storage, curation, interoperability. Technology watch
IT Services	Software, storage, authentication/access, training, support. IT Services can help identify technologies needed by researchers to maximise the value of their research; they can also give advice on how to structure data. Technology watch

Table 2: Roles and Responsibilities

62 Lyon, L., 2007. *Dealing with Data: Roles, Rights, Responsibilities and Relationships. Consultancy Report.* UKOLN. Retrieved at http://opus.bath.ac.uk/412/

ture for the storage of data, but also training and guidance in best practice for data management and citation. Management play a crucial role in driving cultural change within the institution, articulating and reinforcing its orientation towards open access, both through incentivisation and enforcement, and ensuring that the value of open data is recognised.

128. Support services such as research libraries and IT Services are already experiencing demand from researchers to provide support for the creation of data management plans, archiving of data, finding data, and data citation[63]. Depending on the nature of the data and the resources available, libraries may take responsibility for the curation and archiving of data. If we take the example of the humanities into account, libraries have already taken on responsibility for the curation of digital cultural heritage, which is the data used for research in the digital humanities. In disciplines that produce smaller and more heterogeneous datasets it may be the library that provides the infrastructure to make these data available. At the very least, guidance regarding data management plans, open access infrastructures, the basics of good data management, intellectual property and data appraisal and citation should be provided by the library. Good communication across roles is key and with external networks constitutes best practice.

129. Technology watch is a responsibility still being developed. Librarians, IT Services and Data Scientists in particular must maintain a continuous watch on the evolution of technologies (IT support, systems), formats (metadata and backup files) and management plans. That watch must apply as well to standards, which have a strong impact on the work of these groups of professionals.

New role: data scientist

130. Data science has the potential to become a discipline in its own right[64]. In some respects combining the skills of computer scientist and librarian, data scientists can play an increasingly vital role within disciplinary and interdisciplinary research teams, supporting them in data management, exploitation and reuse. Ideally, however, data scientists should have expertise in the discipline where they are working. The Royal Society has pointed out that, in order to combat competition from private industry for the scarce supply of data scientists, universities will need to put clear career paths in place for data scientists in order to attract and retain them[65]. It will also be necessary to develop more advanced (discipline specific) courses for data scientists to increase their numbers. With the increasing availability of datasets from all disciplines and the development of new standards, data scientists will also need access to CPD in areas such as multidisciplinary data science and standards for interoperability and data citation.

Skills and training

131. The development of skills in data management and data re-use is a key enabler of open data.[66] A solid education drawing on best practice in data management will ensure that research data is both trustworthy and reusable. Many of the practices necessary to make data open are simply good practice for any researcher working with data[67]and should be embedded within postgraduate education, becoming a core academic competency[68] which combines research, information literacy and statistical skills along with subject expertise.

132. To ensure that the value of open data is fully realised, researchers will also need practical skills and support to reuse data and assess its quality. A key enabler of data sharing is the trustworthiness of the data, as well as the reusability of the data. Students and researchers will need to be literate in good data management in order to assess and reuse data.

133. The appropriate training is practical in nature, and

63 Swan, A and Brown, S (2008) *The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs*, page 27. Retrieved from http://eprints.soton.ac.uk/266675/

64 Swan, A and Brown, S (2008)

65 The Royal Society, Science as an Open Enterprise, June 2012, p. 64. Retrieved at http://royalsociety.org/policy/projects/science-public-enterprise/report/

66 Dallmeier-Tiessen S, Darby R, Gitmans K, Lambert S, Suhonen J, Wilson M (2012), *Compilation of Results on Drivers and Barriers and New Opportunities*. Retrieved from http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=Compilation+of+Results+on+Drivers+and+Barriers+and+New+Opportunities

67 White et al. (2013) *Nine simple ways to make it easier to (re)use your data*. PeerJ PrePrints 1:e7v1. Retrieved from http://dx.doi.org/10.7287/peerj.preprints.7v1

68 Pryor & Donelly, *Skilling Up to Do Data: Whose Role, Whose Responsibility, Whose Career?* The International Journal of Digital Curation. ISSN: 1746-8256, 2009, Vol. 4, No. 2, pp. 158-170. See http://dx.doi.org/10.2218/ijdc.v4i2.105

ideally discipline-oriented. This means that it is possible to use basic training, formal training programmes for the different roles, supplemented by specialist training. A basic training could include the aspects of ownership, safety/ethics, responsibilities, (re-)use, archiving and sharing of data.

134. For the leadership and management of the institution, it will be necessary for them to ensure mechanisms are in place so that they receive regular briefings on policy and infrastructural developments in the area of open data to support them in their role.

135. Table 3 is a Table of training needs and routes for skills development. Postgraduates and PhDs have been identified as having different training needs to researchers as this is the point where good data management practices should be embedded and also where awareness of open data is instilled.[69] Established researchers, on the other hand, may be motivated by more immediate needs, such as aligning their data management plan with funder requirements.

Recommendations for supporting roles and responsibilities

136. LERU institutions, their support services, European universities, and researchers should:

137. Embed data management courses within postgraduate training.

138. Create a data management support service and information point.

139. Provide general information and guidance on the topic of open research data.

140. Introduce specific job profiles with career paths for data preparation and quality assurance staff – such staff may be embedded in research groups or hosted in data centres or libraries.

141. Enhance awareness among researchers and the wider community by engaging in information activities and data audits.

142. Involve a broad range of stakeholders in training development and delivery, such as heads of graduate schools with a responsibility for training programmes, the HR department, research librarians, IT directors, accrediting bodies and policy makers.

143. Incorporate data curation into library school education.

144. Recognise and foster data science as a professional discipline and create appropriate career paths for data scientists.

145. Invest in quality (accredited) continuing professional development for both data scientists and librarians.

146. Establish doctoral schools for advanced data management and exploitation to increase numbers of data scientists in different disciplines.

WHO	Postgrad /PHD	Senior Researcher	Librarian	Data Scientist
WHEN	Early stages of postgraduate study	As needed or at beginning of research project/proposal state	CPD for subject librarians/During library education	Discipline-specific academic courses (doctoral)/CPD
WHAT	Basics of data management practice, data citation, data evaluation	Training on discipline-specific data management practices, how to write a data management plan tailored to funder requirements, data reuse skills	Data curation. Some disciplinary-specific e-research methods(TDM)/data collection skills, IT skills	Discipline-specific skills for data management/exploitation/interoperability
HOW	Credited models	Practical training	Accredited CPD/Professional courses	Professional (academic) course and accredited CPD

Table 3: Training needs and routes for skills development

69 Swan, A and Brown, S (2008)

Chapter 7: Recommendations to take forward the Roadmap

147. This chapter brings together all the Recommendations in the LERU Roadmap for Research Data and directs them to the most appropriate audience. Here is the heart of the LERU Roadmap. For LERU institutions to assume a leadership role in research data management, it is important that each LERU member successfully addresses the actions in the Recommendations below. It also enlarges the scope of the discussion in the chapters of the Roadmap by suggesting ways in which the bodies of the EU can help support data-driven science. LERU universities should note the following Recommendations.

Recommendations for institutional policy and decision makers at LERU and other universities

R1. Individual LERU members should explore the formation of an RDM Steering Group or similar which brings together the range of critical institutional stakeholders and provides a forum for planning and operational oversight.

R2. Each LERU member should consider developing an institutional Roadmap for Research Data (if they have not done so already) which sets out the strategic objectives, tasks and actions required for compliance with research funder directives.

R3. Every LERU member should develop and promulgate an institutional data policy which clarifies institutional roles and responsibilities for RDM to all stakeholders in the RDM process. Such data policies could be founded on the principles identified in this LERU Roadmap.

R4. Cost modelling for research data management is still in its infancy, although good work has been undertaken to date and is identified in this Roadmap.

R5. Create a data management support service and information point.

R6. Introduce specific job profiles with career paths for data preparation and quality assurance staff – such staff may be embedded in research groups or hosted in data centres or libraries.

R7. Recognise and foster data science as a professional discipline and create appropriate career paths for data scientists.

R8. Institutions should act on the development of formal policies for promoting and rewarding those generating and sharing data of use to the scientific community.

R9. LERU universities will work together to compare their experiences as they develop and implement RDM policies.

R10. LERU will continue to inform policy at the EU level and beyond on the basis of its members' expertise and experience.

Recommendations for all those involved in the curation of research data

R11. Research data should usefully be placed into the framework of the ODE Data Pyramid, to support work on the description and curation of research data.

R12. Documentation and metadata requirements should be identified from the start of any project and considered throughout the lifecycle of the data. Librarians, IT support staff and researchers should work together to identify best practice in metadata formats for describing research data.

R13. Researchers and institutional support staff should work together to identify best practice for the description and citation of research data.

R14. Metadata should comply with existing standards for the content. Chosen formats should preferably support machine-to-machine interoperability. Interoperability between existing CRIS-systems, data curation systems and discovery tools is crucial. Adhering to standards will provide a basis for this.

R15. Research data infrastructure needs to offer a generic framework for a wide variety of research processes and outputs to create, process and share data.

Recommendations for researchers and their institutions

R16. Further work should be done by the LERU research community to achieve consensus, for all the categories of data outlined in the ODE Data Publication Pyramid, on which type of data can be made available for sharing and re-use, or made open, and which cannot.

R17. It is recommended that institutions and researchers work together to clarify what is expected of re-

searchers when citing data. Publication of 'basic principles', general guidelines and examples to help researchers in how to cite data are very common and helpful. This information should include links to existing demands from funding agencies, publishers and data centers, specific to the different disciplines.

R18. It is important to identify the owner of the data: the researcher, funder or institution.

R19. To express the terms of re-use of datasets, it is advisable to use a suitable licence. Researchers should be made aware of this.

R20. Practical support for researchers should be organized. Legal experts, for example the university's legal office, should be prepared to give advice on relevant contracts, regulations and legislation, in addition to information on the website and presented in training sessions.

R21. Embed credited data management courses within postgraduate training.

R22. Enhance awareness among researchers and the wider community by engaging in information activities and data audits.

R23. Involve a broad range of stakeholders in training development and delivery, such as heads of graduate schools with a responsibility for training programmes, the HR department, research librarians, IT directors, accrediting bodies and policy makers.

R24. Incorporate data curation into library school education.

R25. Invest in quality (accredited) continuing professional development for both data scientists and librarians.

R26. Research funders increasingly require data management plans as a part of the submission of project proposals and the LERU research community needs to take note of this development.

Recommendations for LERU Members and the LERU CIO community

R27. LERU universities could exchange information on costs, using the tools and information identified in the Roadmap, to build up a knowledgebase to inform their development.

R28. Were LERU universities to band together to construct shared data management services, the modelling of costings for such a shared service would be an innovative development and help to determine whether there were cost benefits for LERU members in a collaborative approach.

R29. The possibility of collaboration between LERU universities in research data management should be explored, not simply in the area of costings, but also in the areas of the collection and curation of research data. It is important to make sure that services are not unnecessarily duplicated at regional, national or international levels.

R30. The CIO community in LERU has an important role to play in collecting and sharing information on costing for research data management and could act as a focus of best practice in this area.

R31. Provide general information and guidance on the topic of open research data.

R32. Establish doctoral schools for advanced data management and exploitation to increase numbers of data scientists in different disciplines.

R33. LERU Members should engage at international level to build and collect evidence and advocate for the value of open access to research data.

R34. LERU Members should foster a debate amongst stakeholders and disciplines around data sharing.

R35. LERU Members should develop and clearly articulate incentives for researchers to make their data open and should ask funders to do the same.

R36. LERU Members should promote best practice in data management, citation and interoperability to increase the visibility of data and to strengthen the credibility of scientific publications.

R37. A portfolio of tools for an institutional research data infrastructure that fills the gaps in existing external research infrastructures should be develped at LERU universities, e.g. a 'long tail' data repository and a data catalogue supporting re-use and open data. To save costs the sharing of services should be considered.

R38. LERU universitites should establish an asset register of local, central, shared and outsourced research data facilities, allowing continuous monitoring of the efficiency of operations and a strategic approach to capital planning for the 'digital estate'.

R39. To minimize the time researchers have to spend on technical and administrative processes, LERU universities should organize the institutional research data workforce of local IT Officers, subject librarians, central system administrators and research facilitators into a coherent support service.

Recommendations for the bodies of the EU

R40. The EU should encourage and support national stakeholders (governments, funders, universities) to develop research data management policies. Collaboration and dialogue between these bodies should help to align the policies produced by the differing stakeholders.

R41. The EU should engage with universities and facilitate pan-European approaches to the issue of research data management in the context of the European Research Area.

R42. EU funding programmes can help support an approach to data-driven science by introducing funding opportunities for European universities to help them engage with this agenda.

R43. The area of skills development is particularly pressing and EU funding rounds, such as Horizon 2020, should call for proposals to bridge the gaps identified in this Roadmap.

R44. The benefits of Text and Data Mining are being recognised by researchers. To enable secure Text and Data Mining, there need to be revisions, with a Fair Dealing Exception, to the EU Copyright and Database Directives.

About LERU

LERU was founded in 2002 as an association of research-intensive universities sharing the values of high-quality teaching in an environment of internationally competitive research. The League is committed to: education through an awareness of the frontiers of human understanding; the creation of new knowledge through basic research, which is the ultimate source of innovation in society; the promotion of research across a broad front, which creates a unique capacity to reconfigure activities in response to new opportunities and problems. The purpose of the League is to advocate these values, to influence policy in Europe and to develop best practice through mutual exchange of experience.

LERU publications

LERU publishes its views on research and higher education in several types of publications, including position papers, advice papers, briefing papers and notes.

Advice papers provide targeted, practical and detailed analyses of research and higher education matters. They anticipate developing or respond to ongoing issues of concern across a broad area of policy matters or research topics. Advice papers usually provide concrete recommendations for action to certain stakeholders at European, national or other levels.

LERU publications are freely available in print and online at www.leru.org.

University of Amsterdam

Universitat de Barcelona

University of Cambridge

University of Edinburgh

University of Freiburg

Université de Genève

Universität Heidelberg

University of Helsinki

Universiteit Leiden

KU Leuven

Imperial College London

University College London

Lund University

University of Milan

Ludwig-Maximilians-Universität München

University of Oxford

Pierre & Marie Curie University

Université Paris-Sud

University of Strasbourg

Utrecht University

University of Zurich

LEAGUE OF EUROPEAN RESEARCH UNIVERSITIES

LERU Office

Minderbroederstraat 8
B-3000 Leuven
Belgium

tel +32 16 32 99 71
fax +32 16 32 99 68

www.leru.org
info@leru.org