LIBER QUARTERLY
The Journal of the Association of European Research Libraries
Vol. 27, no. 1 (2017) 1–22 | e-ISSN: 2213-056X

Fostering Effective Data Management Practices at Leiden University

Peter Verhaar

Leiden University, The Netherlands
p.a.f.verhaar@library.leidenuniv.nl, orcid.org/0000-0002-8469-6804

Fieke Schoots

Leiden University, The Netherlands
s.p.schoots@library.leidenuniv.nl

Laurents Sesink

Leiden University, The Netherlands
l.b.j.sesink@library.leidenuniv.nl

Floor Frederiks

Leiden University, The Netherlands
f.frederiks@BB.leidenuniv.nl

Abstract

At Leiden University, it is increasingly recognised that effective data management forms an integral component of responsible research. To actively promote the stewardship of all the research data that are produced at Leiden University, a comprehensive, institution-wide programme was launched in 2015, which centrally aims to encourage its researchers to carefully plan the temporal storage, long-term preservation and potential

Uopen Journals | http://liberquarterly.eu/ | DOI: 10.18352/lq.10185

reuse of their data. This programme, which is managed centrally by the Department of Academic Affairs, and which receives important contributions from academic staff, from Leiden University Libraries, and from the University's central ICT organisation, basically consists of three parts. Firstly, a basic central policy has been formulated, containing clear guidelines for activities before, during and after research projects. The central aim of this institutional policy is to ensure that all Leiden-based research projects can effectively comply with the most common requirements stipulated by funding agencies, academic publishers, the Dutch standard evaluation protocol and the European data protection directive. As a second part of the data management programme, faculties have organised workshops and meetings, concentrating on the rationale and on the technical and organisational practicalities of effective data management in order to bring about a discipline-specific protocol. Data librarians employed by Leiden University Libraries have developed educational materials and provide training for PhDs in the principles and benefits of good data management. Thirdly, to ensure that scholars can genuinely make a reasoned selection among the many tools that are currently available, a central catalogue was developed which lists and characterises the most relevant data management services. The catalogue currently provides information about, amongst many other aspects, the organisations behind these services, the main academic disciplines which are targeted and the accepted file formats and metadata formats. The various aspects of these facilities have been classified using terminology provided by conceptual models developed by the UKDA, ANDS and the DCC. Using Leiden University's policy guidelines as criteria, the overall suitability of each service has also been evaluated. Leiden University's data management programme has a total duration of three years, and its basic objective is to offer a comprehensive form of support, in which the data management policy which is propagated centrally is complemented by various forms of assistance which ought to make it easier for scholars to adhere to this policy. The catalogue of data management services also aims to bolster the implementation of an adequate technical infrastructure, as the qualitative evaluations of the services enable policy-makers and developers to quickly establish gaps or other shortcomings within existing facilities.

Key Words: data management; open science; research support; digital scholarship

1. Introduction

Stakeholders across the entire domain of scholarly publishing increasingly recognise that it is of crucial importance to ensure that research data can be curated carefully and responsibly, from the moment when they are created through to their dissemination. While well-considered data management strategies enable researchers to find and to understand their own data whenever they are needed, the various activities that aim to enhance the findability and accessibility of data can also produce clear benefits for the scholarly community at large. Individual research projects rarely exhaust the full potential of the data sets they have created or collected, and these resources often continue to be of value in follow-up investigations, conducted either by the researchers themselves or by others. Different studies may interestingly exploit the data in new resourceful ways, using perspectives that were not envisaged when the data were originally created (see, e.g., European Commission, 2016a; European Union, 2010; OECD, 2007; OECD, 2015; RDA Europe, 2014). It is frequently assumed, moreover, that, when data are made publicly available, this eventually furthers the transparency and the integrity of scholarly work (Lyon, 2016). Well-documented open data can allow peers to replicate the analyses that have been performed, and to verify the conclusions that were drawn from these analyses. Based on the conviction that the public availability of research data can function as a powerful catalyst for the generation of new knowledge, many funding agencies, academic publishers and national and international governments actively stimulate researchers to transfer their data sets into the public domain.[1]

2. Requirements Formulated by Funders, Publishers and Universities

Whereas more and more researchers acknowledge the importance of data sharing, data management generally poses a number of important challenges. Funding agencies, publishers and academic institutions have collectively formulated a confounding number of requirements, and for researchers, it can consequently be difficult to develop a data management strategy which is fully compliant with all of these different demands. As for the management of their data, researchers generally lack a complete overview of the technical facilities that are available, especially when there is no data sharing tradition

within their field. Given the fact that data management tools continue to evolve almost incessantly, it is often difficult to select the most effective data management system. Research support offices at academic institutions usually attempt to offer some guidance in this field, but they have to take into account not only the heterogeneity of research projects, but also the diversity of data management tools.[2] Next to the fact that different disciplines often have different needs with respect to data management, the more specific needs of individual research projects can also vary along with their phase in the overall research process (Whyte and Pryor, 2011). For institutions that do not offer any local facilities in the field of data management, it can be challenging, in short, to formulate appropriate recommendations for the adoption of specific external data management solutions.

Various academic institutions have adopted policies to promote the stewardship of research data, but the challenges that have been listed often complicate an effective implementation of such policies. To create support for the central goals of the data management regulations, policies typically need to be accompanied by a range of additional activities. This paper discusses a number of such activities, drawing from experiences acquired during a comprehensive, institution-wide programme which has been launched in 2015 at Leiden University in the Netherlands. The central objective of this programme is to encourage all researchers who are affiliated with Leiden University to carefully plan the temporal storage and the long-term availability of their data. One important part of this initiative was the development of a clear data management policy. The principles which are laid down in this policy are basically an amalgamation of the main requirements that have been stipulated by funders, publishers and other stakeholders. The work on the data management policy had been preceded by a thorough examination and comparison of these existing requirements.

The analysis focused, more specifically, on the guidelines that were developed for the data management pilot of NWO, the Dutch research council, the pilots on "Research and Innovation Actions" and "Innovation Actions" that are conducted as part of the Horizon2020 programme of the European Commission and in the funding instruments defined by ZonMW, a Dutch agency which predominantly funds research in the life sciences. The three sets of guidelines all specify that grant applications must include a separate paragraph which clearly explains how the data to be produced will be made

available for reuse. NWO expects researchers to describe the facilities they intend to use during and after the research, as well as the concrete measures they intend to take to make these data findable and understandable (NWO, 2016a; 2016b).[3] Grant applications for the Horizon2020 programme must similarly include an outline of a data management plan which takes into account research data quality, sharing and security (European Commission, 2016b). Importantly, the H2020 guidelines pertain to the data that support publications only, and not to the complete collection of raw data. In all the guidelines that have been considered, the initial data management paragraph needs to be expanded into a full data management plan (DMP) when the application is successful. This DMP needs to contain detailed information about the data management environments that will be used, both during and after the completion of the research project. The EC and NWO both suggest that researchers, during their projects, must make use of environments which allow their users to restrict the access to confidential data or to data about identifiable individuals. Researchers whose projects are funded in Horizon2020 are asked to deposit their data in combination with all associated metadata and, if applicable, the software tools that are necessary to replicate and to validate the claims that are based on these data. Importantly, funders recommend the use of trusted and certified data repositories after the completion of the project, where possible.

In addition to the requirements of funders, there are many other types of requirements which Dutch researchers must take into account. In the Netherlands, the Standard Evaluation Protocol, which was developed by NWO and the Dutch academic associations KNAW and VSNU, stresses that researchers need to take measures to secure the integrity of their research. Among other aspects, research integrity covers the manner in which a research project "deals with and stores raw and processed data" (VSNU, KNAW and NWO, 2016). The *Dutch Code of Conduct*, which was formulated by VSNU, the Association of Dutch Universities, emphasises, among other regulations, that researchers need to ensure that their findings are verifiable (VSNU, 2014, p. 8).[4] It stresses that raw data need to be stored for a period of at least ten years. These data ought to be archived in such a way that they can be made accessible quickly and with as little efforts as possible at the request of other researchers. Additionally, in the Netherlands, the general law which protects personal data has been expanded on 1 January 2016, with an article focusing on data leaks (Rijksoverheid, 2016). This law clearly

has implications for the level of security of personal information as well. Next to funders and organisations such as the VSNU and KNAW, academic publishers also form important stakeholders within the field of data management. At present, a growing number of publishers have adopted a Data Availability Policy (DAP), which implies an obligation to make all the data that are referred to in publications available. A condition for publishing in Nature, for instance, is that authors need to "make materials, data, code, and associated protocols promptly available to readers without undue qualifications" (Nature, 2016).[5] In Nature's DAP, it is explained that large data sets should preferably be made accessible through structured public repositories, or, alternatively, through unstructured repositories such as figShare or Dryad. PLOS has defined very similar rules for authors: it asks its authors to provide unrestricted access to all the data that underlie the findings which are discussed in articles, and recommends the use of repositories that demonstrably comply with sustainability and quality criteria, such as those formulated by the *Research Libraries Group* or the international *Data Seal of Approval* (PLOS, 2016).

3. Leiden University's Research Data Policy

Since it can be difficult and time-consuming for researchers to learn about the intricacies of all of these requirements, Leiden's data policy presented itself, more or less, as an abridged version of these numerous guidelines and regulations. An important aim of Leiden's institutional policy was to ensure that all Leiden-based research projects can effectively comply with the most common requirements stipulated by funding agencies, academic publishers and the Dutch standard evaluation protocol. Leiden University's data policy was first formulated in March 2015 and it was subsequently adopted in April 2016 (Universiteit Leiden, 2016). Using the existing requirements and guidelines as a general framework, the policy guidelines divide the research process into three distinct stages: before, during and after the project. In agreement with the requirements of the main funding agencies, Leiden University expects each research project to develop a data management plan before the actual commencement of the project. During the research project, all data needs to be stored safely. This entails, more specifically, an obligation to ensure the integrity, the availability and, if applicable, the confidentiality of the data. After the completion of the research project, researchers must ensure that all

the data sets which are amenable to reuse need to be preserved for a period of at least 10 years. During this period, the data needs to be managed according to the FAIR data principles.[6] The policy also expects researchers to curate their data in combination with the associated metadata, all other documentation which is necessary for this re-use and, potentially, with software tools needed for the analysis of these data. In step with the demands of most funders and publishers, researchers are obliged to deposit their data in a trusted digital repository. When data are stored in a repository which has been awarded the data seal of approval (DSA), scholars can safely assume that their data will be preserved for the longer term.[7] To allow others to cite data sets, it is also advisable to deposit these in an archive which assigns persistent identifiers.

4. Stakeholder Involvement

The formulation of a clear data management policy is obviously a first crucial step, but having a policy in itself is not sufficient. Making sure that researchers actually accept and follow this policy is equally important. To heighten the support for the policy guidelines, much emphasis has been placed on actively involving stakeholders and on making sure that these stakeholders can actively influence important aspects of the policy. The programme was coordinated centrally by the department of Academic Affairs, the office which supports the University's Executive Board. It was directed by a large steering committee consisting of the University's rector magnificus as the business executive, the deans of the various faculties, a representative from University Libraries Leiden and a representative from the ISSC, the central ICT centre. The project team mostly consisted of employees from University Libraries Leiden and from the ISSC. Because it was recognised that data management guidelines can have consequences for virtually all organisational units within the university, it was deemed necessary to collect much feedback at an early stage, from as many different stakeholders as possible.[8] To actively engage academic staff, a number of senior researchers from each faculty were asked to act as data management pioneers. The main task of these pioneers was to organise pilot projects, in which researchers could be familiarised with data management skills, and which could shed some light on the main obstacles within the various research institutes. Together with the research support staff from the faculties, these data management pioneers also formed a sounding board, which convened on a bimonthly or trimonthly basis to

discuss the results and the issues of the different pilot projects. This sounding board has also commented on preliminary drafts of the central data management policy. The programme's inclusive organisational structure effectively allowed the various stakeholders to express their needs and their concerns. Although the different organisational units occasionally had different needs and expectations, there was a general consensus about the importance and the complexity of data management. By and large, the data management programme was welcomed as a means to improve the quality of the research, and as an important step in the integration of research output.

Importantly, Leiden's central data management policy consists of a number of principles only. Since Leiden University is a comprehensive university with seven faculties in the arts, sciences and social sciences, it did not seem feasible to impose a single set of regulations. The policy principally states a number of general goals, which still need to be translated into concrete instructions at the level of individual faculties and research institutes. This approach, in which a decentralised implementation process is combined with a centralised coordination, is based on the assumption that any consensus about the concrete ways in which data is to be managed can only be reached at the level of individual academic disciplines, and not at the level of a university in its entirety. While the institutional policy states, for instance, that particular types of data may need to be made available for re-use, the decision on what needs to be stored exactly and on what can be discarded is to be taken by researchers within specific disciplines. It had been decided, for this reason, that the initial phase, in which the central policy was formulated, needed to be followed by an implementation phase, with a duration of three years, in which the individual faculties develop discipline-specific data management protocols.

5. A Catalogue of Data Management Services

Leiden University's policy was defined as part of an institution-wide programme which centrally aimed to remove some of the crucial difficulties that researchers may encounter while planning for their data management. As it was clear that researchers often lack the skills and the knowledge to select data management facilities, a number of activities had been planned to develop the knowledge of researchers around data management and to offer

detailed, specific information on the tools that are available. To ensure that scholars can genuinely make a reasoned selection among the many tools that are currently available, a central catalogue was developed which lists and which characterises the most relevant data management services. The aim of the catalogue was to make it easier for researchers to select an appropriate data management service. This registry is clearly not the first nor the only one of its nature. One notable example of a similar initiative is the re3data registry, which was launched in 2012.[9] Arguably, the list of data archives which have been awarded the *Data Seal of Approval* likewise serves as a catalogue of services from which research teams can choose (Data Seal of Approval, 2009). One potential shortcoming of existing registries is that they do not directly clarify the circumstances under which the various data management services can be useful for specific research projects. Furthermore, they generally focus on one specific type of service or on one specific stage in the research data life cycle. The catalogue that was developed at Leiden University built on the data that had already been compiled for existing catalogues. It aimed to make it easier for researchers to compare the different services on the basis of a number of criteria, and to relate the properties of these services to the data management policy of Leiden University.

For the purpose of the catalogue, an extensive description model was developed, largely based on the requirements by funders and publishers that were discussed above. To be able to offer advice to scholars who are searching for appropriate data management tools, it is necessary to have reliable information not only about the requirements of external parties, but also about the ability of software tools to meet such demands. The functionalities that are offered by data management tools and the concrete ways in which these systems can differ can be described effectively using terminology borrowed from a number of existing models of data curation processes and of research data in general. The Data Curation Centre, for example, has created a useful model of the various activities that can be performed by data archives. The model provides a "graphical high-level overview of the lifecycle stages required for successful curation" (Higgins, 2012).[10] The Australian National Data Service has similarly developed a model that can be used to classify data facilities. The model postulates that data can occur in three distinct domains. In the private domain, researchers manage their own research data on their own storage devices. When scholars collaborate in a team, it is often useful to move the data to a shared domain, in which all team members

can access the data. Data sets which are finished and which can be cited in publications need to be transferred to a public domain. In this latter domain, research data are mostly available in combination with metadata and with other documentation (Treloar, Groenewegen, & Harboe-Ree, 2007). Because data facilities are often designed for a specific stage in the research process, it was also necessary to make use of a model which effectively represents the various phases in the scholarly lifecycle. The UKDA data lifecycle model proved useful in this context, since it has a data-centric approach, concentrating specifically on activities performed on data.[11] Data have also been classified using the typology that was proposed by Reilly, Schallier, Schrimpf, Smit and Wilkinson (2011) in their conceptualisation of the "research data pyramid". In this model, a distinction is made between (1) raw data and data sets; (2) data collections and structured databases; (3) processed data and data representations and (4) publications with data (Reilly et al., 2011). Data facilities can be compared by considering the degree to which they provide support for different types of data, and for the various activities in the field of data curation.

The description model that was used for Leiden University's catalogue of data management services currently consists of 59 fields in total, and they are listed in Table 1. Using this description model, it became possible to collect information about the various data management facilities in a highly systematic manner. Amongst many other aspects, the information sheets offer information about the organisations behind these services, the relevant legal aspects, the main academic disciplines which are targeted and the accepted file formats and metadata formats. In addition to this, the various aspects of these facilities have been classified using concepts and terminology from the four models that have been described above.[12] The catalogue currently provides information about ca. 50 local, national and international services. The focus was predominantly on services that are already used or are likely to be used by researchers from Leiden University.

6. Suitability of Data Management Services

The information sheets listing aspects of data management tools were created mainly to enable researchers at Leiden University to make a well-considered choice when planning the management and the storage of their data.

Table 1: Description model for data management services.

General information	Url of the tool, description, organisation, type of service, usage and appreciation, support organisation
Context	Stage in the research project, position within the research process, domain, type of data, data curation activities, data classification
Administrative information	Funding, depositor agreement, user agreement policy, intellectual property, data curation strategy
Target groups	Faculty, primary target group, secondary target group
Classification of the service	Availability, integrity, confidentiality
Formats	Accepted metadata formats, accepted content types, accepted preferred formats, accepted file formats
Storage	Maximum size of deposits, version management, quality control
Access	Access requirements, tools or interfaces for access, (persistent) identifiers
Preservation	Long term guarantees, compliancy with international standards for trusted repositories, preservation strategy
Costs	Costs for storage, costs for access, costs for preservation
Special conditions	Agreements with Leiden University, risks

To make this process of selection easier, the various services have also been evaluated in a qualitative sense, on the basis of the criteria that are listed within Leiden University's data management policy. The services that predominantly aid activities during active research projects have been evaluated using the following criteria:

- The service must take measures to ensure the integrity of the data. Data integrity means that the accuracy and consistency of data is maintained and assured over their entire life-cycle.
- Within data management services, it must be possible to make data available to others. This does not necessary imply full open access. It may also entail a provision through which data can be made available exclusively to funders or peer reviewers, for example.
- When the public availability of data would violate existing privacy or copyright protection laws and regulations, the service must be capable of securing the confidentiality of these data.
- To ensure that research data can be findable and intelligible, it must be possible to store data in combination with metadata and other relevant documentation.

Data archives that aim to secure the long-term preservation of research data must satisfy not only the criteria above, but also the following requirements:

- The service must assign persistent identifiers to the data sets that have been deposited.
- The organisation that is responsible for the repository must guarantee that the data set can be preserved for a period of minimally 10 years.
- The data repository must have been granted a relevant certification (DSA, DIN, RAC/ISO 16363).
- The data repository must accept data formats whose longevity can be guaranteed.
- The organisation that is responsible for the data repository must have formulated a mission statement, which explicitly mentions the ambition to preserve data for the longer term, and which also expresses a clear vision on the ways in which the funding for their services can be sustained.

Using these requirements, an assessment was made of the suitability of the listed data management services. A distinction was made between their suitability before, during and after the research project. When the service met all of the requirements, it was considered suitable. If only some of the requirements were met, the service was considered to be partly suitable. If the service adhered to none of these principles, it was deemed unsuitable. When a service clearly lacked any functionalities for one of the three general stages in research process, the evaluation for this particular phase was set to "not applicable".

The evaluations that were added can be useful for a variety of reasons. Most pertinently, they can be beneficent to researchers who aim to select a data management tool that is adequately in step with Leiden University's data management guidelines. In addition to this, these qualitative assessments can also be useful for the organisations that are responsible for these tools. The information sheets can effectively help developers to identify lacunae or other shortcomings. The evaluations of the services that are available locally at Leiden University usefully indicated, for instance, that none of these are suitable for use after the completion of research projects. To ensure the long term preservation of data, researchers clearly need to make use of services

that have been developed elsewhere. The analyses additionally made clear that only two of the services that have been developed locally are actually suitable for use during research projects. This information is relevant to policy makers responsible for investment in new research infrastructures.

Fig. 1: Complete list of services with an indication of their suitability, measured against the criteria mentioned in Leiden University's data management policy. A green check sign means that the service meets all requirements, an orange question mark indicates that the service meets some of the requirements and a red cross implies that the service meets none of the requirements. The codes "B", "D" and "A", which are used in this overview, stand for "Before", "During" and "After", respectively. They refer to the three research phases which are distinguished in Leiden University's policy. The information that is shown on the website is dynamic in nature. This image shows the information that was known to the project team at a specific moment in time.

7. Website About Data Management Services

The information that had been accumulated about the various data management facilities are presented on a website that is publicly accessible.[13] As can be seen in Figure 1, the results of these qualitative assessments of the data management services are also shown on the project's public website using colour codes and icons. Users of the site can easily navigate through the list by making selections based, for instance, on the general phases in the research project or on the academic disciplines which can make use of the service. On the basis of these browsing facilities, researchers can easily find information on services with very specific characteristics. Importantly, the four theoretical models that have been used can also be used to navigate through the contents of the site. By clicking on one of the nodes in the visual representation of UKDA's data lifecycle model (Figure 2), visitors of the site can quickly identify tools which primarily help to analyse data or to re-use data (Figure 3), for example.

Fig. 2: Navigation based on research data lifecycle model.

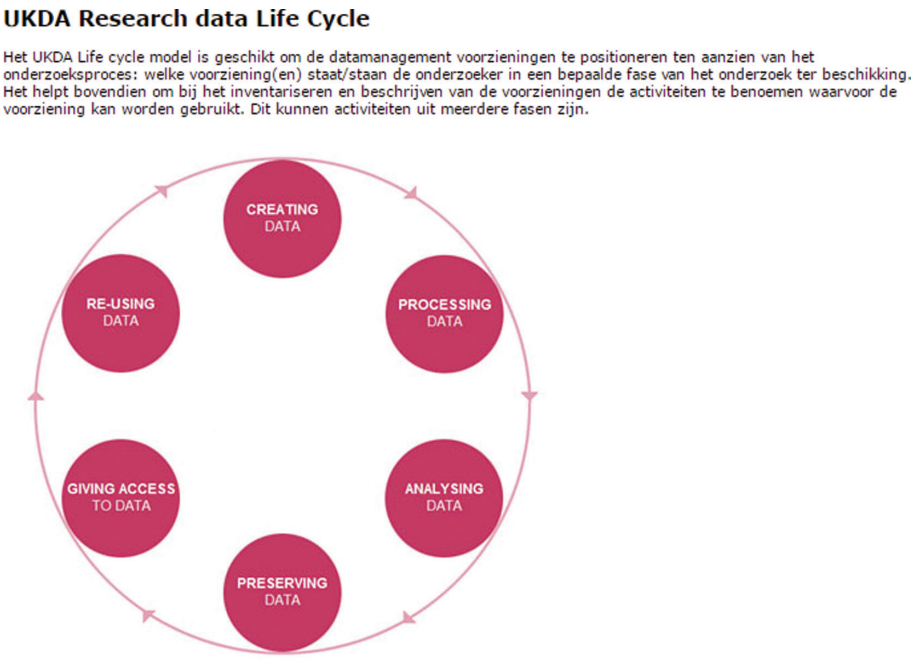
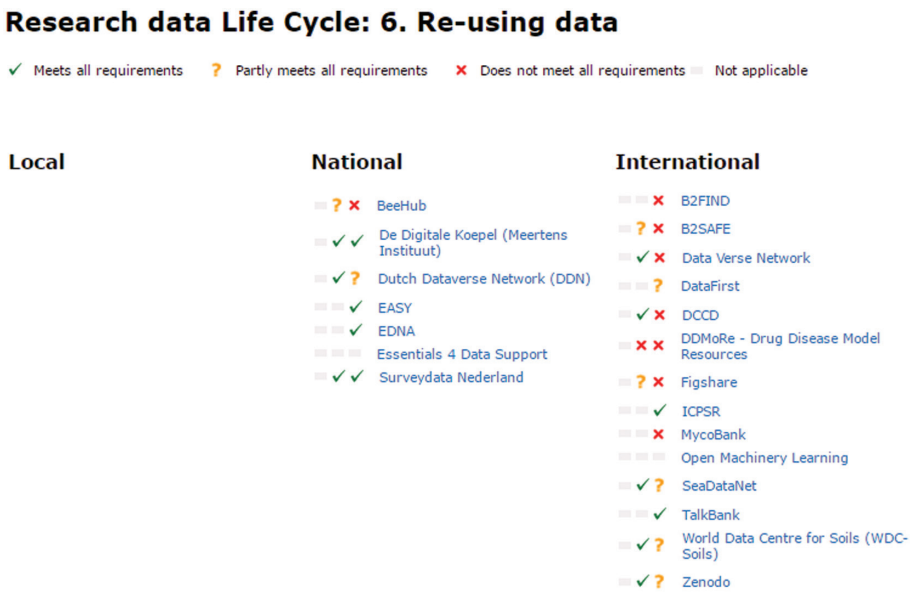


Fig. 3: Navigation based on research data lifecycle model. For information about the meaning of the colour codes and abbreviations that are used in this overview, see the caption of Figure 1.



The website is still under development. Some of the information that is presented on the site still needs to be verified and, if necessary, updated. As data management facilities continue to evolve, the task of verifying and editing information appears to be of a structural nature. In the coming years, we aim to explore the further development of the catalogue with partners within and outside of the university, and to organise the content management of the site in a sustainable way.

8. Training and Education

Next to the catalogue of data management services, researchers can also be supported by the two data librarians employed by the Centre for Digital Scholarship (CDS), which is located physically and organisationally within the University library. The CDS has developed educational materials and it organises two types of training sessions. Firstly, the CDS offers data

management training on demand for groups of PhD students, postdocs and other researchers of Leiden University. Such training sessions are of an interactive nature, and they are tailor-made in close collaboration with research institutes. Such sessions usually concentrate on the sticks and carrots of data management planning, on the principles of good data management and on best practices in individual research fields. Participants write a DMP according to the University's template. Next to these discipline-specific forms of training, the CDS also organises general introductory courses on data management planning, which focus, amongst other topics, on the rationale of Leiden University policies, funder requirements, the technical and organisational practicalities of effective data management, principles of secure storage, journal policies and data archives. These courses are open to all employees of the Leiden University. In addition, the CDS informs researchers on internal and external requirements concerning data management and gives advice on the appropriate sections in their research proposals and on their full data management plans.

9. Conclusion

The experiences at Leiden University suggest that the success and the impact of data management policies depend, to a large extent, on the availability of ancillary activities that can promote the acceptance of such data management guidelines. Universities can create support for data policies by actively engaging with all relevant stakeholders during the full policy making process, and by combining centralised general principles with local implementation processes, in which researchers can develop discipline-specific data management protocols. Next to this, it is also vital to make sure that researchers have access to detailed and up-to-date information about existing data management facilities, and that they can follow both custom-made and generic courses on data management. The ultimate aim of Leiden University's data management programme is to offer a comprehensive form of support, in which the data management policy that is propagated centrally is complemented by a range of activities, which essentially follow a bottom-up approach. This broad range of activities should ultimately help researchers to find their way in a very dynamic and a highly complicated area of expertise.

At the time of writing, Leiden University's data management programme is still ongoing, and it is difficult, for this reason, to formulate firm or

conclusive statements about the effectiveness of the programme. The fact that the various meetings and training sessions that are organised attract increasingly large audiences appears to be a clear indication of a growing awareness of the importance of data management. It has been decided, moreover, that the usage and the appreciation of the various services that are being developed need to be monitored closely after the end of the programme. Leiden's data management team is currently considering some of the ways in which the diffusion of data management skills can be measured. One suggestion is to develop a central facility for the storage of all new data management plans. It would also be very useful if all researchers could be encouraged to register their data deposits centrally, potentially in the Leiden University's CRIS. At present, all faculties are translating the central general data management principles to discipline-specific protocols, and some of the measures that have been taken during these implementation processes underscore the notion that researchers acknowledge the value of responsible data management practices. At the Institute of Psychology, for instance, attending a data management training has become mandatory for all PhD students. The Leiden Academic Centre for Drug Research similarly made data management training a compulsory part of the PhD Education and Supervision program. At this same institute, PhD students can only defend their thesis when they have made their data available in the exact same way that is described in their data management plan.

Virtually all research-intensive universities face the necessity to provide their academic staff with sound advice on how to manage and to curate their research data, and many institutions can benefit from a qualitative comparison of the main services that are currently available in this context. The utility of the catalogue that was developed during Leiden's data management programme is clearly not limited to one institution. As was stressed above, the descriptive model underlying this catalogue was derived from the main requirements that have been formulated by funding agencies and by publishers, and these evidently apply equally to scholars at other universities. A number of universities in the Netherlands have already indicated that it can be beneficial to develop the catalogue collectively, and to expand it into a resource that can eventually be used by all Dutch universities.[14] At the same time, it can be useful to explore the feasibility of an international version of the catalogue.

Although it has been mentioned at several points in this text that a careful data management can help scholars to satisfy the various requirements of funders and of publishers, the focus should clearly not be too narrowly on these requirements. The various rules and regulations ultimately serve as a means to an end, and the various data management policies are generally formulated to encourage researchers to actually reap the many benefits that can emanate from the public availability of copious quantities of data. The tools that can be used to analyse and to visualise big data collections become increasingly sophisticated, and a growing number of scholars have begun to experiment with the manifold benefits that may ensue from such data-intensive forms of research. As measuring devices, digital research instruments and statistical software packages continue to generate data sets, it becomes increasingly important for data scientists and other researchers to develop effective methods to exploit these data and to extract relevant and significant patterns from these petabytes of resources. As the result of such exertions, studies may begin to answer traditional questions differently or they may even begin to ask questions that were previously impracticable or inconceivable. Crucially, such data-intensive forms of research depend on the presence of reliable scholarly infrastructure and on a judicial and well-considered use of tools, which can effectively guarantee the continued availability, findability and usability of data.

References

Akers, K.G. (2014). Going beyond data management planning: Comprehensive research data services. *College & Research Libraries News*, *75*(8), 435–436. Retrieved January 18, 2017, from http://crln.acrl.org/content/75/8/435.full.pdf+html.

Data Seal of Approval (2009). *Community*. Retrieved November 25, 2016, from http://www.datasealofapproval.org/en/community/.

European Commission (2016a). *The European cloud initiative. Digital single market*. Retrieved November 12, 2016, from https://ec.europa.eu/digital-single-market/en/%20european-cloud-initiative.

European Commission (2016b). *H2020 programme. Guidelines on FAIR data management in Horizon 2020*. Version 3.0. Retrieved November 12, 2016, from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

European Union (2010). *Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on scientific data*. Retrieved

November 18, 2016, from http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707.

Higgins, S. (2012). The lifecycle of data management. In G. Pryor (Ed.), *Managing research data* (pp. 17–46). London: Facet Publishing.

Krier, L., & Strasser, C.A. (2012). *Data management for libraries: A LITA guide.* Atlanta: ALA TechSource.

LERU Research Data Working Group (2013). *LERU Roadmap for research data*. Retrieved December 8, 2016, from http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf.

Lyon, L. (2016). Transparency: The emerging third dimension of open science and open data. *LIBER Quarterly*, *25*(4), 153–171. Retrieved November 18, 2016, from https://doi.org/10.18352/lq.10113.

Nature (2016). *Data availability statements and data citations policy: Guidance for authors*. Retrieved November 20, from http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf.

NWO (2016a). *Data management protocol*. Retrieved November 12, 2016, from http://www.nwo.nl/en/policies/open+science/data+management.

NWO (2016b). *Data management section*. Retrieved November 12, 2016, from http://www.nwo.nl/en/policies/open+science/data+management+chapter.

OECD (2007). *OECD principles and guidelines for access to research data from public funding*. Paris: Organisation for Economic Co-operation and Development. Retrieved November 12, 2016, from http://www.oecd.org/sti/sci-tech/38500813.pdf.

OECD (2015). *Making Open Science a Reality.* In OECD Science, technology and industry policy papers, No. 25. Paris: OECD Publishing. Retrieved November 20, 2016, from https://doi.org/10.1787/5jrs2f963zs1-en.

PLOS (2016). *Data availability*. Retrieved November 20, 2016, from http://journals.plos.org/plosone/s/data-availability.

Pryor, G., Jones, S., & Whyte, A. (2014). *Delivering research data management services: Fundamentals of good practice*. Edinburgh: DCC.

RDA Europe (2014). *The data harvest report - Sharing data for knowledge, jobs and growth*. Retrieved December 18, 2016, from https://www.rd-alliance.org/data-harvest-report-sharing-data-knowledge-jobs-and-growth.html.

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). *Report on integration of data and publications. Opportunities for Data Exchange (ODE)*. Retrieved November 12, 2016, from http://libereurope.eu/wp-content/uploads/ODE-Report OnIntegrationOfDataAndPublication.pdf.

Rijksoverheid (2016). *Wet bescherming persoonsgegevens*. Retrieved November 25, 2016, from http://wetten.overheid.nl/BWBR0011468/2016-01-01.

Treloar, A., Groenewegen, D., & Harboe-Ree, C. (2007). The data curation continuum. Managing data objects in institutional repositories. *D-Lib Magazine 13*(9–10), n.p. Retrieved November 12, 2016, from http://www.dlib.org/dlib/september07/treloar/09treloar.html. https://doi.org/10.1045/september2007-treloar.

Universiteit Leiden (2016). *Research data management regulations Leiden University*. Retrieved November 12, 2016, from http://regulations.leiden.edu/research/research-data-management-regulations-leiden-university.html.

VSNU (2014). *The Netherlands code of conduct for academic practice. Principles of good academic teaching and research*. Retrieved November 12, 2016, from http://www.vsnu.nl/files/documenten/Domeinen/Onderzoek/The_Netherlands_Code%20of_Conduct_for_Academic_Practice_2004_(version2014).pdf.

VSNU, KNAW, & NWO (2016). *Standard evaluation protocol 2015–2021. Protocol for research assessments in the Netherlands – Amended version*. Retrieved November 12, 2016, from https://www.knaw.nl/nl/actueel/publicaties/standard-evaluation-protocol-2015-2021.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data 3*, 160018, n.p. Retrieved November 12, 2016, from https://doi.org/10.1038/sdata.2016.18.

Whyte, A., & Pryor, G. (2011). Open science in practice: Researcher perspectives and participation. *International Journal of Digital Curation*, *6*(1), 199–213. Retrieved November 19, 2016, from https://doi.org/10.2218/ijdc.v6i1.182.

Notes

[1] The LERU Roadmap for Research Data emphasises, among similar lines, that "main point of making data openly available is so that it may be reused for new purposes" (LERU Research Data Working Group, 2013, p. 12)

[2] Good overviews of the various ways in which research libraries can meet the needs of researchers in the field of data management can be found in Pryor, Jones and Whyte, 2014, Krier and Strasser, 2014 and Akers, 2014.

[3] NWO also states, notably, that the costs for data management can be entered in the project budget.

[4] Peers should be able to assess whether or not the data and the conclusions drawn from these data comply with "relevant standards (for instance of quality or reliability)". Researchers should guard "[t]he quality of data collection, data input, data storage and data processing", and they should ideally document the various steps that have been taken in a study, in resources such as lab journals, project reports and minutes of important meetings. VSNU emphasises that "[c]onduct is verifiable when it is possible for others to assess whether it complies with relevant standards (for instance of quality or reliability)". (VSNU, 2014, p. 8)

[5] Nature states that "authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications".

[6] FAIR stands for Findable, Accessible, Interoperable and Reusable. See, for example, Wilkinson et al., 2016. As Leiden University's data policy was formulated before the FAIR acronym became more widespread, it does not explicitly use the acronym. It emphases, nonetheless, that data need to remain "findable, accessible, comprehensible and reusable".

[7] Next to the DSA, there are two additional levels of certification. As described in the *European Framework for Audit and Certification of Digital Repositories* (http://www.trusteddigitalrepository.eu/Trusted%20Digital%20Repository.html), data archives can either perform a self-audit on the basis of the ISO 16363 or DIN 31644, or they can be subjected to an external audit on the basis of the same protocols.

[8] The approach that was followed by Leiden University was in step with the main recommendations from the LERU Research Data Working Group. The LERU Roadmap stresses that "advocacy needs to occur at every level within the institution and beyond". When developing data management policies, institutions need to engage actively with the appropriate stakeholders "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" (LERU Research Data Working Group, 2013, pp. 11–12).

[9] Registry of research data repositories, (http://www.re3data.org/). Re3data was developed by the Berlin School of Library and Information Science at the Humboldt-Universität zu Berlin, the Library and Information Services department (LIS) of the GFZ German Research Centre for Geosciences, the KIT Library at the Karlsruhe Institute of Technology (KIT) and the Libraries of the Purdue University.

[10] The model (see Higgins, 2012) distinguishes seven activities: (1) creating or receiving data; (2) appraisal and selection; (3) ingest of data; (4) preservation actions; (5) storage; (6) facilitating access, use & reuse and; (7) transforming data. The final activity which is mentioned can entail processes through which obsolete data formats are converted into more sustainable formats, or processes in which unused or redundant data are deleted.

[11] The model lists six stages: (1) creating data; (2) processing data; (3) analysing data; (4) preserving data; (5) giving access to data and (6) re-using data. In all of these stages, the nature of the data can be different.

[12] The description model that was developed at Leiden University displays many similarities with the list of criteria which were formulated for the data repository comparison tool which was developed by MIT, see https://libraries.mit.edu/data-management/share/find-repository/.

[13] The URL of this website is <https://vre.leidenuniv.nl/vre/lrd/>

[14] This activity is coordinated by the Workgroup Facilities and data infrastructure of the National Coordination Point Research Data Management, see https://www.surf.nl/en/lcrdm/issues/facilities-and-data-infrastructure.