Data Management in Practice

Results and Evaluation
January 2018

DEff Danmarks Elektroniske Fag- og Forskningsbibliotek

Contributors:
Aalborg University Library: Karsten Kryger Hansen
Danish National Archives: Christina Guldfeldt Madsen, Anne Sofie Fink, Bodil Stenvig
DTU Library: Falco Jonas Hüser and Paula Maria Martinez Lavanchy
Roskilde University Library: Søren Møller, Stine Vejlebo, Sacha Zurcher
Royal Danish Library, Aarhus: Tony Brian Albers, Bjarne Andersen, Katrine Hofmann Gasser, Knud Åge Hansen, Bolette Ammitzbøll Jurik, Filip Kruse, Jesper Boserup Thestrup The Royal Danish Library, Copenhagen: Anders Sparre Conrad and Michael Svendsen University Library of Southern Denmark: Janni Brødbæk and Asger Væring Larsen

ISBN: 978-87-7507-415-0
DOI: 10.7146/aul.243.174

Aarhus, January 2018
Denmark's Electronic Research Library (DEFF)

Table of Contents

1: Executive Summary and recommendations 6
1.1: Recommendations based on the findings of the project 7
1.2: Practical outcomes of the project 8
2: Introduction 9
2.1: The six themes and their objectives 10
2.2: The outline of this report 11
3: Humanities Cases 12
3.1: LARM Case: Infrastructure for media studies (Royal Danish Library, Aarhus) 12
3.1.1: Description of the case 12
3.1.2: Course of events 13
3.1.3: Experiences 13
3.1.2: Data Management Planning 14
3.1.3: Data capture, storage, and documentation (basic metadata) 14
3.1.4: Data identification, citation and discovery 15
3.1.5: Select and deposit for long-term preservation 15
3.1.6: Sustainability 15
3.2: Netlab Case: Internet studies (Royal Danish Library, Aarhus) 16
3.2.1: Description of the case 16
3.2.2: Data Management Planning 17
3.2.3: Data capture, storage, and documentation (basic metadata) 17
3.2.4: Data identification, citation and discovery 18
3.2.5: Select and deposit for long-term preservation 18
3.2.6: Sustainability 19
3.3: LARM, Netlab and data management planning 19
3.4: Model Agreement on Data Management in Cooperative Research Projects – or to put it briefly, a legal reflection of a Data Management Plan 20
3.5: Kierkegaard Case (Royal Danish Library, Copenhagen) 21
3.5.1: Common perspectives – 50 years of access for continued research 21
3.5.2: Using DMPonline with researchers 22
3.5.3: The Kierkegaard Case 23
4: Social Science Cases 26

4.1: Local Elections Surveys (University of Southern Denmark, The Danish National Archives) 26
4.1.1: Description of the case 26
4.1.2: Progress during the project 26
4.1.3: Questions raised include: 27
4.2: CALPIU - Centre for Cultural and Linguistic Practices in the International University (Roskilde University Library) 28
4.2.1: Description of the case 28
4.2.2: Data Management Planning 29
4.2.3: Data storage, capture and documentation 29
4.2.4: Select and deposit for long-term preservation 29
4.2.5: Data identification, citation and discovery 30
4.2.6: Sustainability 30
5: Health Science Cases 31
5.1: The GALAXY project (University of Southern Denmark) 31
5.1.1: Description of the case 31
5.1.2: Data Management Planning 32
5.1.3: Data storage, capture and documentation 32
5.1.4: Select and deposit for long-term preservation 32
5.1.5: Data identification, citation and discovery 33
5.1.6: Sustainability 33
6: Science Cases 34
6.1: CAFF terrestrial CBMP - Technical support to operationalize parts of the Conservation of Arctic Flora and Fauna (CAFF) terrestrial Circumpolar Biodiversity Monitoring Programme (CBMP) (Roskilde University Library) 34
6.1.1: Description of the case 34
6.1.2: Data Management Planning 35
6.1.3: Data storage, capture and documentation 35
6.1.4: Select and deposit for long-term preservation 35
6.1.5: Data identification, citation and discovery 36
6.1.6: Sustainability 36
7: Technological Sciences Cases 37
7.1: DTU Wind Energy: Meteorological data (Technical University of Denmark) 37
7.1.1: Description of the case 37
7.1.2: Goals 38
7.1.3: Approach 38
7.1.4: Expected outcomes 39

7.1.5: Results and conclusions 39
7.2: DTU Space: Geomagnetic data (Technical University of Denmark) 43
7.2.1: Description of the case 43
7.2.2: Goals 44
7.2.3: Approach 44
7.2.4: Requirements 45
7.2.5: Outcomes 45
7.3: The Kepler Case 48
7.3.1: Infrastructure perspectives 50
7.3.2: Conclusion for Royal Danish Library, Copenhagen, cases 51
8: Common conclusions of the cases 52
9: Thematic areas 53
9.1: Data Management Planning 53
9.1.1. Danish DMPonline installation established and available 53
9.1.2. DMPonline adapted to Danish researchers and Danish conditions 53
9.2: PURE and research data management 55
9.2.1. Registration in PURE 55
9.2.3. PURE as a repository for documents related to data management 56
9.2.4. PURE for registering data management activities 56
9.3: Data capture, storage and documentation 57
9.4: Data identification, citation and discovery 57
9.4.1. Background and success criteria 57
9.4.2 Revised success criteria 58
9.4.3 Linking and visualisations 59
9.4.4 DataCite – recommended as exchange format 59
9.4.5 Results 60
9.5: Select and deposit for long-term preservation 60
9.5.1: Success Criteria 60
9.5.2: Experiences 61
9.5.3: Success Criteria 1: Documentation and Best Practice 61
9.5.4: Success Criteria 2: Data from all cases preserved in the long term 62
9.6: Training and marketing toolkits 65
9.6.1: The lifecycle and DMP challenge – a detour 65
9.6.2: Data lifecycle as a flow – back on track 66
9.6.3: DataFlowToolkit for training 69
9.7: Sustainability 69

9.7.1: Theme 1: Data Management Planning 69
9.7.2: Theme 2: Data capture 70
9.7.3: Theme 3: Data identification 70
9.7.4: Theme 4: Long-term preservation 71
9.7.5: Theme 5: Marketing and training 71
9.7.6: Cases 71
9.8: Conclusion of the thematic section 72
10: Evaluation of the project by the international experts 74
10.1: About the two experts 75
11: Conclusions 76
12: References 79

1: Executive Summary and recommendations

The Data Management in Practice project produced both tangible and practical results. Firstly, hitherto missing parts of a Danish data infrastructure able to cover the entire lifecycle of research data were established. Secondly, perhaps less tangible, but still practical, it was proved that research libraries and archives have a vital role to play in research data management and that their expertise can be regarded as an integrated part of national Danish solutions.

As a consequence of the project, the combined Danish research data infrastructure now contains a national DMPonline server (the template used on the server meets the demands inherent in H2020) and the Library Open Access Repository (LOAR), which offers repository facilities for open data.

Libraries and archives already operate services such as the Netarchive, The Danish Data Archive and The DeIC National Cultural Heritage Cluster, as well as taking part in earlier data management projects. Participation in this project has further qualified the staff working with the services.

We can describe the structure of the project as a hybrid middle path between a purely case-based project with individual institutions each working on their own sub-projects, and a thematic project with institutions working within one or more broad themes. The structure chosen contained six themes: Data Management Planning; Data capture, storage and documentation; Data identification, citation and discovery; Select and deposit for long-term preservation; Training and marketing toolkits; and Sustainability. Each of the participating institutions worked on specific cases, such as research projects, data collections etc., with the cases also designed to cover the entire data lifecycle. Each case should relate to the themes in order to draw conclusions on both a more case-specific and a more general theme-specific level. The cases covered the main academic fields of Humanities, Social Sciences, Science and Technology.

The case-based approach succeeded to a great extent in providing the framework for the development of practical solutions in close cooperation with researchers within the cases' specific fields. Furthermore, it showed that the infrastructures of the academic fields – IT-based and 'human' infrastructures in the form of networked services for guidance, legal advice, training, etc. – across the different academic fields are of great importance. This stresses the need for political and institutional awareness of the issue of data management and the need for provision of the necessary funding.

The cases also demonstrate that close cooperation between departments at the individual university regarding IT, legal and other forms of research support is important. The supporting services have to be designed and developed in close co-operation with researchers in order to ensure their relevance.

The thematic approach demonstrated on the one hand that the themes were well-suited to providing practical insight into the processes of research data management. On the other hand, it showed that the solutions developed have to be diverse in order to suit the specific needs and demands of the research projects.

1.1: Recommendations based on the findings of the project

Based on the results of the work on the project cases and themes, we can deduct some recommendations regarding the development and operation of systems and services for data management.

Firstly, we recommend that the different institutions develop and adopt policies regarding research data management. As an example, such policies will help the researcher to know which data the institution thinks must be preserved. These policies must cover the entire data lifecycle from fund application, through active research and sharing of research data, and finally to preservation in data repositories. The policies can be based on regulations or established practices in the research communities.

These policies must of course meet local institutional requirements, but we recommend that the policies are flexible and, in terms of design, take account of national policies adopted by other Danish research institutions, as well as corresponding to international standards. This is necessary in order to ensure that local policies do not create extra workloads for researchers if the project spans several institutions. If requirements vary significantly or are institutionally unique, they will result in extra workload for researchers. This can be exemplified by the research data management policy adopted by the University of Southern Denmark, which was largely based on the policy of the Technical University of Denmark (DTU). This enabled the two institutions to minimize the risk of extra workload.

Furthermore, the policies must be flexible and adjustable in order to handle different data formats, different quantities of data and changes in these during the project in question. Our cases contain examples of such variations and thus stress the need for flexibility in order to meet researchers' demands.

Due to the legal and technical complexities of research data management, a precondition for handling the problems at a national level is close cooperation between the various research libraries, the Royal Danish Library and the Danish National Archives. We recommend that these institutions cooperate closely in order to ensure that regulations, different requirements and services are coordinated in order to ensure agile workflows based on the needs of researchers. As a minimum, we recommend that an organizational framework for cooperation – a network – is established between staff members from the above institutions in order to ensure permanent exchange of information. We suggest that the network already established during this project could be developed as the core of a larger network in the future.

This kind of cooperation is also necessary at institutional level. The development of local services covering the entire lifecycle of research data will require cooperation between the legal department, IT department, library and other relevant parties in terms of research support. We therefore recommend that each research institution organizes an internal network in order to ensure a comprehensive flow of knowledge and experience.

Finally, we recommend provision of the necessary funding for the operation of national services in order to ensure the continuation of national services – such as those we present in the next section – and the prevention of individual duplication by different institutions, which would risk being less cost-efficient compared to national services. Furthermore, these national services must avoid "Danish" solutions in favour of a more international perspective.

1.2: Practical outcomes of the project

The project not only generated experience with data management, but it also resulted in active services and applied knowhow.

As a direct result of the project, a national installation of DMPonline has been established[1]. It is operated by Danish e-Infrastructure Cooperation (DeIC) and, as a project funded by Denmark's Electronic Research Library (DEFF), it is meant to ensure that the services are used by as many researchers as possible. As part of the project, a national template for a Data Management Plan has been developed and can be tested and used by researchers via the national DMPonline server. The server thereby gives access to several templates.

Several of the cases involved in the project showed that, as expected, legal issues cause problems for researchers in sharing and accessing research data. In order to help researchers avoid such problems, a format for a contract regarding data management was developed. The contract is meant to be a flexible standard contract which researchers can use when sharing data with external partners. The contract was designed in cooperation with Aarhus University[2].

In order to train researchers and other employees, a presentation was developed of the "Dataflow Toolkit"[3] to demonstrate how Data Management can affect a given research project and help researchers in the process of designing their workflow. The toolkit is publicly available in a beta version.

Finally, as a result of the experience gained during the project, the Royal Danish Library has established a data repository in order to give researchers the possibility of sharing research data and to make data based on the collections of the institution available as research objects to researchers. The service is called Library Open Access Repository (LOAR)[4].

[1] https://dmponline.deic.dk/, accessed 10262017
[2] https://www.deic.dk/da/news/2017-08-15/modelaftale (in Danish only), accessed 10262017
[3] https://dataflowtoolkit.dk/, accessed 10262017
[4] http://www.statsbiblioteket.dk/datarepo/ , accessed 10262017 and https://loar.statsbiblioteket.dk/xmlui/ , accessed 10262017

2: Introduction

This report presents the results of the Data Management i Praksis (DMiP) project (in English: Data Management in Practice). The project was funded by Denmark's Electronic Research Library (DEFF)[5], the National Danish Archives and the participating main Danish libraries. The project began in March 2015 and was completed in October 2017. The project had a budget of DKK 9 million (Approx. 1.2 million Euro). The following partners participated in the project: Aalborg University Library[6], Danish e-Infrastructure Cooperation (DeIC)[7], DTU Library[8], Roskilde University Library[9], The Danish National Archives[10], Royal Danish Library[11] and The University Library of Southern Denmark[12].

The project was assessed by international experts in order to ensure that the project had achieved the objectives described in the application. The evaluation took place at a conference on 30th March 2017.

The general objective of the project was to establish a Danish setup covering all aspects of the lifecycle of research data: from application, through the research phase, and finally to the dissemination of results and sharing of research data. The setup was to be based on researchers' demands, and the suggestions and results of the project were to be at an international level. The project should also demonstrate that research libraries have a role to play regarding research data. Furthermore, the project should ensure development of competences at the libraries, which can then be used in the future process of managing research data[13].

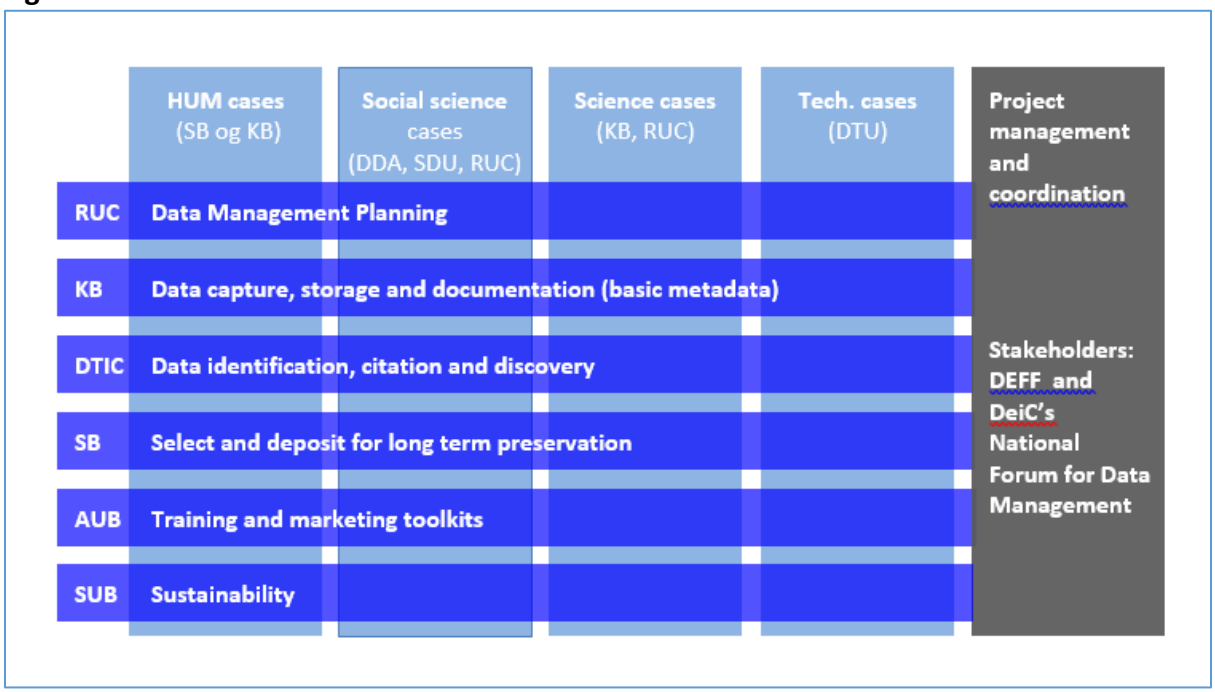
In 2015, the Danish e-infrastructure Cooperation (DeIC) established the Nationalt Forum for Forskningsdata Management (National Forum on Research Data Management)[14]. This was done in close cooperation with Danish universities, research libraries, national libraries and the Danish National Archives. One objective of the DMiP project was to evaluate how the results of the DMiP project were reflected in the work of the National Forum[15].

---

[5] http://www.deff.dk/english/ , accessed 10262017
[6] http://www.en.aub.aau.dk/, accessed 10262017
[7] https://www.deic.dk/?language=en, accessed 10262017
[8] http://www.bibliotek.dtu.dk/english, accessed 10262017
[9] https://rub.ruc.dk/en/, accessed 10262017
[10] https://www.sa.dk/en/, accessed 10262017
[11] In 2015, two national libraries participated in the project: The Royal Library and The State and University Library. In autumn 2016, the two libraries were merged into The Royal Danish Library. Since the activities related to the project were undertaken in Aarhus and Copenhagen respectively, we refer to the specific sections of the library as The Royal Danish Library, Aarhus, and The Royal Danish Library, Copenhagen. When a given activity includes the former name of the institution, we use the unchanged name. You can find the websites here: http://www.kb.dk/en/index.html and http://www.statsbiblioteket.dk/ (both accessed on 10262017)
[12] http://www.sdu.dk/en/%20bibliotek, accessed 10262017
[13] Andersen 2014, 7. See Appendix 1.
[14] https://www.deic.dk/datamanagement/DM_forum, accessed 10262017
[15] Andersen 2014, 7. See Appendix 1.

The project was divided into a number of cases and into a series of broad themes. The cases would cover all academic fields. The themes were defined to cover the entire lifespan of research data. Each case should, if possible, draw conclusions related to each theme and thereby make it possible to draw more general conclusions. You can see how the different cases are related to the different themes in Appendix 2. Each institution involved participated in a least one case and each institution was responsible for a specific topic. Each theme had its specific objectives[16]. A description of the cases and the themes can be found in sections 3-7 and 8.

Figure 2.1: The distribution of involved institutions in terms of cases and themes.



2.1: The six themes and their objectives

The six themes were: Data Management Planning; Data capture, storage and documentation; Data identification, citation and discovery; Select and deposit for long-term preservation; Training and marketing toolkits; and finally Sustainability.

Data Management Planning was chosen as a general topic in order to ensure the establishment of an infrastructure able to address researchers' requirements in the initial phases of a given research project. The objective of the theme was twofold: Firstly, to establish a national installation of DMPonline[17]; and secondly, to develop a Danish national template that matches Danish conditions based on the experience gained from working with Danish researchers[18]. Roskilde University Library coordinated the activities of this theme.

[16] Andersen 2014, 7-9. See Appendix 1.
[17] https://dmponline.deic.dk/, accessed 10262017
[18] Andersen 2014, 7. See Appendix 1.

The theme of data capture, storage and documentation addresses the problems researchers have regarding the lack of infrastructure that makes it possible to capture and store research data during the research phase. The objectives of the theme were that the project should establish repositories and implement technologies able to capture data[19]. The Royal Danish Library, Copenhagen, coordinated the activities of this theme.

The theme of data identification, citation and discovery relates to issues concerning the dissemination and sharing of research data. The objectives in this case are to ensure that the metadata of shared research data contains unique IDs of the datasets. Furthermore, the project must ensure that the data is indexed in a national search engine or index[20]. DTU Library coordinated this theme.

The theme of select and deposit for long term preservation relates to issues that researchers will have to respond to in order to store and preserve research data for prolonged periods. The objectives of the project were to create relevant documentation on how to select and store data and to ensure that research data was stored in a format that would make long-term storage and sharing possible[21]. The Royal Danish Library, Aarhus, coordinated the activities in this theme.

The theme of training and marketing toolkits relates to the fact that data management for many staff members is a new working area. In order to establish services, it is necessary to train the staff. The objectives of the theme were that the project should design training tools and reach out to different groups of researchers[22]. Aalborg University Library coordinated this theme.

The theme of sustainability addressed problems which arise when a certain service is maintained beyond its original project period. The theme is aimed at developing and describing different business models which could ensure the maintenance of relevant services[23]. The University Library of Southern Denmark coordinated the activities in this theme.

2.2: The outline of this report

The outline is a reflection of the structure of the project. The first two sections consist of an executive summary and a general introduction. This is followed by a section with more detailed descriptions of the cases, before a thematic section which summarizes the findings from the cases at a thematic level. The last sections consist of a presentation of the evaluation performed by the international experts and a concluding section.

[19] Andersen 2014, 7-8. See Appendix 1. Page 7-8, see Appendix 1.
[20] Andersen 2014, 7. See Appendix 1.
[21] Andersen 2014, 7. See Appendix 1.
[22] Andersen 2014, 7. See Appendix 1.
[23] Andersen 2014, 7-9. See Appendix 1.

3: Humanities Cases

3.1: LARM Case: Infrastructure for media studies (Royal Danish Library, Aarhus)

Table: 3.1: The data involved in the LARM project:

Project:	**LARM**
Type of data:	Data on the server consists of recorded TV and radio programmes, various written programmes and metadata. Existing data formats: Video, audio, PDF, JSON files Data stored: annotations made by researchers Data format created via the case: JSON
Amount of data which should be stored in a repository	Approximately 1,200 files containing 12,000 annotations. In total 6 MB.

3.1.1: Description of the case

As a project, LARM[24] (LydArkiv for Radiomedier = sound archive for radio media) was active from 2010 to 2014. It involved 11 different institutions: The University of Copenhagen, Roskilde University, The University of Southern Denmark, Aalborg University, Aarhus University, The Kolding School of Design, The Royal School of Library and Information Science, Danish e-Infrastructure Cooperation, the Danish national broadcasting service (DR), The Media Museum Brandts and the Royal Danish Library, Aarhus. LARM has established a database with radio programmes from 1925 up to the present day. LARM's main purpose is to establish a digital archive and research infrastructure with the appropriate tools to enable researchers and students to use radio programmes as a source for research, primarily within the humanities (Kruse, Thestrup, 2014, p. 324)[25]. Following the end of the project, LARM was handed over to the national consortium, DIGHUMLAB[26]. LARM is in the process of becoming DIGHUMLAB theme 2b, research in audio-visual media. A need has already been expressed for management of research data and preservation. Of special importance is the relation between primary research data and objects (e.g. from cultural heritage) and new research data.

The various research projects work with radio data through the interface LARM.fm. This enables researchers to work with the material and enrich it with new metadata, correct existing metadata, annotate and sort radio programmes into file structures to facilitate search and discovery.

[24] http://www.larm.fm, accessed 10262017
[25] Kruse and Thestrup, 2014.
[26] https://dighumlab.org, accessed 10262017

The aim of this case is to establish an overview of the new data and to establish a solution for long-term preservation of data (location and format). Furthermore, the aim is that data should be reusable, e.g. in mediestream.dk and for other research projects.

As a participant in the Data Management in Practice project the Royal Library, Aarhus will harvest and preserve the user-created data in LARM and the metadata necessary to link this data to the original programmes and other archived materials.

The participants are: the Royal Danish Library, Aarhus; DIGHUMLAB; Iben Have, Associate Professor at AU; Janne Nielsen, Assistant Professor at AU; and Chaos Insights[27] (running the LARM.fm platform).

3.1.2: Course of events

The LARM case has followed three tracks:

- The more general track of developing the DMP template
- The case-specific track of analyzing and deciding which data to preserve – existing data harvested in the old system or existing data harvested in the new system in order to offer a future harvesting facility.
- The technical track: analyzing and deciding on harvesting and preservation of data (formats, etc.)

3.1.3: Experiences

LARM has been useful as a test case for the development of the DMP template of the Royal Danish Library, Aarhus. This has contributed to the improvement of the template, the clarification of the organization of the future service and demonstrated the need for legal and IT competence.

Several meetings with researchers have been held in order to clarify and work out the content and design of the case. The replacement of researchers during the course of the project and the uncertain status and organizational affiliation of LARM have emphasized the importance of Data Management Planning.

The technical set-up of the LARM.fm platform asks the researcher to confirm that all data entered into the system can be shared using a CC0 licence[28]. The annotations (data) entered are not provided with identification of the author, which can be regarded as a shortcoming in the metadata. The annotations have an internal ID and sometimes refer to an internal ID for the radio/TV programme with timestamps. The metadata for the radio/TV programme includes a "DOMS-ID", which can be used to locate the programme in mediestream.dk. The data can be harvested as JSON-data. It includes the annotations, the metadata for annotations and the metadata for the associated programme.

[27] http://www.chaosinsight.com/, accessed 10262017
[28] https://creativecommons.org/share-your-work/public-domain/cc0/, accessed 01092018

We had assumed that it would be possible to share the data with reference to the CC0 licence. It turned out, however, that some of the annotations contain sensitive data, which means that this is not possible. The metadata for the associated programme contains programme descriptions from Ritzau and Datameter, which is copyrighted data and cannot be shared freely either.

We looked into the possibilities concerning anonymization, but though natural language processing has come very far and we can anonymize free text to some degree, we cannot guarantee that no sensitive data will remain after an automatic anonymization process. At present, we recommend preserving the data with access restricted to individual researchers by application. We are also looking into the possibilities of sharing data with the same restrictions as the original programmes and metadata. They also include both copyrighted and sensitive data, but they can be accessed by university staff and on location at the Royal Danish Library[29]. The Data Management Plan appears in Appendix 3.

3.1.2: Data Management Planning

The Data Management plan for the LARM case was developed together with the template. This means that it was developed over several iterations with involvement from both DIGHUMLAB and the Royal Danish Library, Aarhus. The final DMP is attached, see Appendix 3. Read more about developing the DMP template in section 3.3: LARM, Netlab and data management planning.

LARM has been useful as a test case for the development of The Royal Danish Library, Aarhus, DMP template. This has contributed to the improvement of the template, the clarification of the organization of the future service and demonstrated the need for legal and IT competence.

Several meetings with researchers have been held in order to clarify and work out the content and design of the case. The replacement of researchers during the course of the project and the uncertain status and organizational affiliation of LARM have emphasized the importance of Data Management Planning.

3.1.3: Data capture, storage, and documentation (basic metadata)

The original cultural heritage data presented to researchers in the LARM.fm platform are radio and TV broadcasts with metadata. The practical data capture was done by harvesting all “assets” with annotations from the LARM.fm platform. In this case, an asset is a representation of all metadata associated with a single audio or video file.

The technical set-up of the LARM.fm platform asks the researcher to confirm that all data entered into the system can be shared using a CC0 license. The annotations (data) entered

[29] Further reading: Jensen Kreutzfeldt, Michelsen and Svendsen, 2015, and Andersen, Larsen and Tøgersen, 2013.

are not provided with identification of the author, which can be regarded as a shortcoming in the metadata. The annotations do have an internal ID and sometimes refer to an internal ID for the radio/TV programme with timestamps. The metadata for the radio/TV programme includes a "DOMS-ID", which can be used to locate the programme in mediestream.dk. The data can be harvested as JSON-data. It includes the annotations, the metadata for annotations and the metadata for the associated programme.

3.1.4: Data identification, citation and discovery

We had assumed that it would be possible to share the data with reference to the CC0 licence. It turned out, however, that some of the annotations contain sensitive data, which means that this is not possible. The metadata for the associated programme contains programme descriptions from Ritzau and Datameter, which is copyrighted data and cannot be shared freely either.

We had expected to get a persistent ID along with citation and discovery platform for free when depositing the data into an open access archive. This has not been possible. We now recommend using a restricted access repository and still getting the persistent ID from the repository.

We should note that this is a fairly large, very heterogenic dataset, and we would expect researchers to prefer to cite smaller and more homogenic datasets. The option for the researcher to download and share smaller datasets has been suggested in the Data Management Plan. With a small, well-known dataset, it is possible for the researcher to ensure that there is no "problematic" data in the dataset before preserving and sharing it.

3.1.5: Select and deposit for long-term preservation

For the large dataset, we looked into possibilities concerning anonymization, but though natural language processing has come very far and we can anonymize free text to some degree, we cannot guarantee that no sensitive data will remain after an automatic anonymization process. We recommend preserving the data with access restricted to individual researchers by application. We are also looking into the possibilities of sharing data with the same restrictions as the original programmes and metadata. They also include both copyrighted and sensitive data, but they can be accessed by university staff and students, as well as on location at the Royal Danish Library.

3.1.6: Sustainability

We recommend the data to be preserved in the long term by the Royal Danish Library, Aarhus. A service is being established at the time of writing, see section 9.5.4.2: LCAR: Library Controlled Access Repository. We should note that the data has also been reported to the National Archives for appraisal.

3.2: Netlab Case: Internet studies (Royal Danish Library, Aarhus)

Table: 3.2: The data involved in the Netlab project:

Project:	**Netlab**
Type of data:	Existing data from the Netarchive stored in ARC and WARC files: approx. 770 TB Data format created via the case: CDXJ + bash, python, CSV, JSON
Amount of data generated	Under 5 TB

3.2.1: Description of the case

The project title is: Probing a Nation's Web Domain — the Historical Development of the Danish Web. The first aim is to map the historical development of the entire Danish web, based on the available material in the national Danish web archive, Netarkivet[30].

The second aim is to develop the necessary research infrastructure to conduct such a study, i.e. tools and procedures to handle corpus creation, a variety of analyses, long term storage, documentation, workspace and collaborative working tools.

The main research question to be answered is: What did the Danish web as a whole look like in the past, and how has it developed? A fundamental element in this mapping is to investigate the methodological questions involved in conducting the study, as well as to establish and continuously develop the analytical design.

The third aim is to establish a well-documented corpus of the Danish web from each year. The documentation of the creation of this corpus of the Danish web, year-by-year, is to be kept by Netarkivet, with a view to being used by other scholars in the future. Thus the project has a research, as well as a (double) research infrastructure, dimension.

Finally, in a longer perspective the project intends to be the cornerstone of a European project aimed at mapping and comparing the different national web domains in Europe, possibly within the framework of RESAW[31], a Research Infrastructure for the Study of Archived Web, which is being established as a transnational project aimed at submitting a Horizon 2020 application.

The role of the DMiP project in this case is primarily assistance in data management planning, advice on documentation and metadata, and help with long-term preservation.

[30] http://netarkivet.dk/in-english/, accessed 10262017
[31] http://resaw.eu/, accessed 10262017

3.2.2: Data Management Planning

The first meeting between the Data Management In Practice project and the Netlab Case was a joint meeting with the DeIC National Cultural Heritage Cluster and The Royal Danish Library[32], where all three aims were presented.

One of the issues we discussed was the role of libraries in data management and in particular with regard to long-term preservation. Niels Brügger, Professor, Head of the Centre for Internet Studies and of NetLab/DIGHUMLAB, expected documentation of the creation of corpora and other research data to be kept close to the original data, which is at Netarkivet, with the aim of being used by other scholars in the future. Netarkivet does not offer such a service (yet), but the Royal Danish Library is establishing one, see section 9.5.4.2: LCAR: Library Controlled Access Repository.

Since then, we have held 5 Data Management Planning Meetings and we have attached the final DMP (in Danish), see Appendix 4. See also section *3.3: LARM, Netlab and data management planning.*

3.2.3: Data capture, storage, and documentation (basic metadata)

We defined three levels of corpora descriptions

- Researcher description of corpus (already written)
- Technical description / algorithm (work in progress)
- The machine readable version

Existing data is data from Netarkivet. The owner of this data is the Royal Danish Library. The data contains personal identification information for human subjects, as well as copyrighted material. Access is given to the individual researcher. The amount of data used can be up to 300 TB. Given the nature of the internet, the file formats can in principle be any known file formats. The data is stored in ARC and WARC files and the metadata in Crawllogs.

New research data will be indices of specified corpora from Netarkivet and procedures and descriptions for generating indices, as well as derived research data such as descriptive statistics, procedures and descriptions for generating descriptive statistics.

As part of this project, the aim is to develop the necessary research infrastructure to make a study of a nation's web domain. One challenge is to develop tools, procedures, formats, documentation, etc.

A lot of work has gone into this step and has been documented in the DMP and in Probes — The Cookbook[33], which is a “cookbook” for internet research.

This case changed character while waiting for the Cultural Heritage Cluster to become fully functional. During 2014, a large part of Netarkivet was made searchable through Solr[34]. A lot of web archive statistics could now be generated by the search engine, and new research

[32] http://en.statsbiblioteket.dk/kulturarvscluster, accessed 10262017
[33] See Appendix 5.
[34] https://sbdevel.wordpress.com/net-archive-search/ ,accessed 11012017

data now includes statistical search results, queries and descriptions of how to generate queries and read statistical search results.

3.2.4: Data identification, citation and discovery

Again, we are hoping for a repository solution. If we use the LOAR / LCAR solutions, see sections 9.5.4.1: LOAR: Library Open Access Repository and 9.5.4.2: LCAR: Library Controlled Access Repository, both these repositories supply DataCite DOIs, which makes data searchable through DataCite's services. They also both expose metadata using OAI-PMH, which means the metadata can be harvested by and made searchable through Dansk Data Arkiv[35].

3.2.5: Select and deposit for long-term preservation

One challenge encountered early in this case when working with the DMP is *ownership.* Who owns the data? This is different for data that existed before the start of the project and data that is produced as part of the research project.

In this case, the data that existed before the start of the project is the national Danish web archive, Netarkivet. This data contains personal identification information for human subjects, as well as copyrighted material, and access is restricted to approved research. This also means that we have to be careful about research data generated in the project. If we are sure the research data does not contain sensitive information or copyrighted material, we may share the data. If we are not sure, the data should be just as restricted as the Netarkivet data, or it should be deleted.

The central datasets to preserve are the corpora descriptions. We can preserve and share the 3 types of descriptions identified. This is, however, not exact, as the Danish Web Archive is not static. New data is harvested and material from different archives is occasionally incorporated. This means that rerunning the machine readable algorithm will not always yield the exact same corpus, and thus the statistical analyses performed on the corpus will not yield exactly the same result.

We can instead preserve the resulting corpora "indices". Using the references to the material in the web archive will give us an exact definition. The indices used for analysis are in CDX format. This is, however, not a standardised format and is dependent on current technology. We propose a new corpus definition[36] using the pwid[37], which is a precise and constant web reference.

The indices generated may contain personal identification information or copyrighted material, and they will not be shared. We recommend preserving the indices at the new Royal Danish Library "Controlled" Access Repository, where it is possible to restrict access to only individually approved researchers, see section 9.5.4.2: LCAR: Library Controlled Access *Repository*. Procedures and descriptions for generating indices, as well as derived data, do

[35] https://www.sa.dk/en/services/dda-danish-data-archive/danish-data-archive/, accessed 10162017
[36] Jurik and Zierau 2017.
[37] Zierau, Nyvang and Kromann, 2016.

not contain personal identification information or copyrighted material. These will be preserved and shared using the Royal Danish Library Research Data Repository service, LOAR, see section 9.5.4.1: LOAR: Library Open Access Repository.

3.2.6: Sustainability

We should note that the Netlab case data has been reported to the National Archives for appraisal, and the National Archives declined to preserve the data. The data is to be preserved in the long term by the Royal Danish Library, Aarhus, using the two repository services LOAR and LCAR. The Royal Danish Library is funded by the Danish state, and the risk of the service being terminated is small.

3.3: LARM, Netlab and data management planning

The LARM and NetLab cases were used to develop a local template (see Appendix 6) for a Data Management Plan for researchers who wish to have access to data stored by the Royal Danish Library in Aarhus. The process began by asking researchers already working on projects using data from the radio and television collections and the web archive to fill out an early version of the template. This approach was chosen in order to gain experience on how to work with a Data Management Plan and to ensure that the researchers were presented only with relevant questions. Based on these results, the finished template was designed and uploaded to the test version of DMPonline.

The template can be used by the researcher to manage the entire research workflow and also by the library to prepare specific technical elements of the services. By combining the two elements, the library expects to save time for both researchers and library staff.

In general, the template is based on the Horizon 2020 template combined with some specific technical questions related to the specific collections of the Royal Danish Library. Since the services of the library aim to cover access to archives, analytic tools and online storage of research data, the questions cover the entire lifecycle of the data.

The work soon demonstrated the benefits of systematic data management. For example, one of the respondents experienced problems in completing the template due to several changes taking place in the group of researchers working on the project and also changes to versions of the software used in the project. This information was not readily available to the respondent.

The early versions of the template contained many questions, which gave researchers the impression that they were actually duplicating work in order to fill out the template. Many questions in the template asked for information which the researchers had already provided, compiled in applications for funding of the specific projects. In order to save time and effort for researchers, the template now simply provides the option of referring to the relevant documents, such as applications for funds and other data management plans.

The first versions of the template also asked many technical questions regarding the format of data, rights to use data and rights to share data. The template still addresses these issues, but the number of questions has been reduced and the template is now planned to be part of a flexible, integrated service, i.e. a member of the library staff will assist the researcher in filling out the template.

The cases demonstrated that sharing research data requires a comprehensive legal framework to be built into the template. As the legal status of data will change in the event that a project moves from research in archived data to dissemination of data, the template also needs to address legal problems and assist the researcher and the library in ensuring that relevant legislation is not breached. This issue also forms the basis of the formulation of the Model Agreement on Data Management.

The comprehensive testing, and indeed the whole design process, of the template offers valuable insights into the benefits of data management, the need for a flexible approach with regard to use of the template – and the art of limitation when asking questions in a template.

3.4: Model Agreement on Data Management in Cooperative Research Projects – or to put it briefly, a legal reflection of a Data Management Plan

DeIC (Danish e-Infrastructure Cooperation) has funded the project which is carried out by staff from Aarhus University and The Royal Library. This project is to some extent a consequence of the DMiP project. As the development of the local template for Data Management Planning progressed, it became evident that researchers needed to be introduced to the legal framework as soon as possible in the research process in order to avoid legal problems and the burdens of a double task.

The Model Agreement (see Appendix 7) is a comprehensive document, as its intention is to cover as many variants of potential legal issues related to data management as possible. At the same time, it is flexible and adjustable to fit all areas of research.

The aim of this project is to lighten researchers' administrative workload by providing them with an instrument to clarify legal questions related to their projects' data and results:

- Systematic decision-making on issues of ownership and copyright to research data
- Regulating access to data during cooperative research projects and after their completion
- Regulating publication rights to research results
- - all of the above include variants of researchers' entries and exits in ongoing research projects -
- An overview for researchers of potential issues connected to the use of e-infrastructure not provided by, for example, the researchers' own universities.

The Model Agreement has been evaluated and tested by researchers from Aarhus University. Among the results are that the Model Agreement is a highly relevant instrument, as it can be used both in its entirety and in selected sections relevant to the individual

project. Several researchers preferred the document to be structured in modules instead of one comprehensive document, as this would allow for easier adaptation to the particular research project. A spin-off result of the analysis was valuable insight into researchers' existing practices in handling research data.

During the work on DMPonline, it became obvious that there is a limit to the level of detail in the listing of questions related to legal aspects of data management. Although an independent project, the Model Agreement can also be regarded as a solution to this problem, and as such it represents a valuable extension of the Data Management in Practice project. The report, as well as the Model Agreement, is available at DeIC's homepage[38].

3.5: Kierkegaard Case (Royal Danish Library, Copenhagen)

Table: 3.3: The data involved in the Kierkegaard project:

Project:	**Kierkegaard**
Type of data	The full dataset consists of 212 archival units ("writings"), in turn consisting of 651 TEI files. Data is packaged in BagIt format for machine readability and long-term stewardship.
Amount of data which should be stored in a repository	Amounts to approximately 15 GB

3.5.1: Common perspectives – 50 years of access for continued research

Both the cases covered by Royal Danish Library, Copenhagen, have been researcher-driven based on a strong wish that data must be preserved and curated for continuous active research for many years to come. In both cases, the data is currently being used by international researcher communities. Independent of discipline, this has raised two key issues:

- Technological and institutional sustainability: who can take care of data with a very long-term perspective – a minimum of 50 years – and how to best prepare the data for this?
- How can data be curated in such a way that it remains accessible for continued development and active research?

While the sustainability and preservation issue has institutional, technological and financial implications, the issue of curation for active research points towards dataset structure, standardized files, metadata formats and discipline-specific discovery needs. Both cases have had a special focus on data reuse for research purposes, with this focus being reflected in the work with data itself, as well as supplying the supporting infrastructure. Although the

[38] https://www.deic.dk/da/news/2017-08-15/modelaftale, accessed 11032017

challenges have seemed considerable at times, our approach has been to move as far as we could towards supplying solutions, rather than getting lost in the potential problems.

The emergence of the FAIR principles[39], even if not part of the formal setup of this project and even if we have not systematically evaluated our solutions with regard to FAIR, have to a great extent guided the work in both cases, e.g. in the choice of machine readable formats and documentation, the structure and content of datasets, the supporting infrastructure, as well as requirements for identification and use of established vocabularies for semantic relations.

3.5.2: Using DMPonline with researchers

Neither of the two research data cases covered by the Royal Library, Copehagen has been driven by any mandatory delivery of a Data Management Plan. Thus addressing different stages of the research lifecycle seemed an obvious and agreed starting point for discussing and facilitating the required dialogue concerning management of the research data in question.

At an early stage, the setup of the pilot installation of the Danish DMPonline became a reality in the DMiP project, from which the Royal Library could create a framework for collaborating with the research community and actively involve researchers in trying to solve the requirements of managing the data. The default DCC template embedded in the DMPonline tool was used as generic guidance to address a tangible way of dealing with a 50-year perspective on securing reuse of active research data for future researchers. By performing future-like workshops with the researchers using DMPonline, important questions arose and turned out to be key challenges to be addressed in the curation of data:

- For which research questions might future researchers find this data useful?
- How would they most likely want to see data packaged?
- What documentation is necessary to understand data outside the current context?
- Which search criteria would most likely be used to discover data?

Rather than answering questions from the DMPonline guidelines step-by-step, as a matter of fact more and increasingly detailed questions emerged concerning how best to structure the datasets suitable for long-term preservation in useful formats that would most likely outlive current applications and software. At the same time, scoping towards this activity made the division of work and areas of expertise quite clear from the beginning across the disciplinary skills involved, which consisted of data science, computer science, data stewardship and information and metadata management.

Finally, working with the DMPonline tool when writing up a living and dynamic document like a DMP collaboratively has proved to be an advantage in terms of structuring our communication and timing of outreach to relevant strategic partners in raising awareness of the process and the ongoing work in data management.

[39] Wilkinson et al, 2016

3.5.3: The Kierkegaard Case

Philosopher Søren Kierkegaard's writings were published a few years ago as a new complete edition, *Søren Kierkegaards Skrifter*, consisting of physical books and a web edition in parallel, in principle with identical content and structure. The technological foundation of this edition was the mark-up of the texts in a self-developed format, KN1[40]. The electronic edition provides the text for reading in HTML format, but does not offer the raw marked up data for research purposes. Moreover, sks.dk is read only, thus providing no chance of further editing the texts, a large part of which are commentaries provided by researchers.

As such, sks.dk threatens to become a closed text, which will die with the expected obsolescence of the technical platform. The research team behind the new edition, however, wanted the text to remain alive and open to researchers, also for the provision of new commentaries and annotations. They wanted it to be possible to add new material to the edition for possible new versions of the edition. There was a distinct feeling that nobody would be going to invest in a new corpus of Kierkegaard's texts for at least the next 50 years, meaning that the current data is very valuable and must remain accessible for research.

Our project case was built with this researcher requirement in mind: to preserve the data in a way that would allow access to the raw texts for continued research. This data stewardship challenge was understood to consist of two main dimensions:

- A sustainable archiving solution that could ensure continued access to the files for many years to come
- Bringing the file content onto a standardized format that could serve both preservation purposes and continued active use and development by researchers.

The main task has been to migrate the data from the self-developed KN1 format to a standardized and widely used format. In this case, TEI[41] was the choice, as this connects the Kierkegaard corpus with other literary publication projects at Royal Danish Library, thus potentially enabling a new electronic version, as well as with broad initiatives in the Humanities sector in Denmark. TEI is well supported by the Clarin.dk platform and is indeed used by other major scholarly editions in Denmark, such as the Grundtvig Edition.

About 80% of the migration could be done relatively easily by automatically mapping KN1 mark-up to standard TEI elements and attributes. However, the last 20%, and in particular the critical apparatus, caused problems. An analysis in cooperation with the research centre's staff revealed the reason to be that the original KN1 mark-up was designed very specifically to support the philological principles of the edition. It would not be possible to complete the format migration without taking this into account. We appointed the computer scientist originally behind the edition, who completed the migration to TEI in cooperation with researchers from the Søren Kierkegaard Research Centre, which is associated with Copenhagen University.

[40] Kierkegaard Normalformat 1. Dokumentation: http://sks.dk/red/kn1dok.xml, accessed 11032017
[41] TEI: Text Encoding Initiative: http://www.tei-c.org/index.xml, accessed 11032017

In order to preserve the files and to make them accessible to researchers, we experimented by depositing them into our pilot Dataverse (see below). We created a Dataverse for the complete corpus[42] with four sub-Dataverses for each category of texts in the edition: published works, unpublished works, journals and papers, and letters and dedications. We created a number of example datasets, each representing one work. Each dataset would then contain 3-4 files, typically the text itself, commentary and text explanations, and, if applicable, the introduction. This would constitute a structure that would be very easy to search and to work with for future researchers. Each title would be identified with a DOI, and each file in a dataset would have its own landing page in Dataverse and be citable/referable on its own.

TEI, being a self-contained object model, is known to cause problems for repository builders, since the descriptive metadata is contained in the file itself. In our case, it was straightforward to extract the metadata from the TEI header and to insert it into Dataverse metadata fields. The process could even be scripted if the entire corpus was to be deposited. Problems would arise, though, in the event that a researcher or curator would be editing the metadata in Dataverse. In this case, the changes would not automatically be applied to the TEI headers for the dataset. Thus in the case of TEI, according to the FAIR principles' separation of data and metadata, this could potentially result in inconsistencies due to the partially conflicting object models.

We were not able to provide a production setup of Dataverse in time for archiving the newly created TEI files by the end of February 2017. So we decided to temporarily archive the files at Copenhagen University's Electronic Research Data Archive, ERDA[43]. ERDA is a large file archive with geo-replication and tape storage. We archived all the migrated files, along with the original file formats and all supplementary material from the original edition which had not previously been properly archived.

At a specific time and date, all the original files were retracted from the current server to a safe file server drive. The conversion script was run on the KN1 files and the resulting TEI files packaged in BagIt format and deposited into the ERDA archive. As an extra safety measure, a read-only freeze archive was made, which will be backed up to tape, as well as residing in the online archive. The TEI files were arranged as a new version of the SKS files, alongside the older version. In addition to being part of the BagIt structure, they were stored in a separate zip archive, potentially allowing researchers to download the entire corpus as one file. For added readability, all the TEI files were run through XSLT conversion and made available as HTML files as well.

ERDA is a file archive only and offers neither identification (DOI), metadata, licensing nor discovery services. In order to allow researcher access, a read-only share link has been generated. This will be given to the Research Centre for them to manage access for the time being. In time, this solution will hopefully be replaced with a real data repository deposition

[42] Søren Kirkegaards Skrifter Dataverse: https://ec2-52-50-247-224.eu-west-1.compute.amazonaws.com/dataverse/sks, accessed 11032017

[43] https://erda.dk, accessed 11032017. Requires password.

of the data as properly structured datasets, with search engine and proper AAI and licensing, including the possibility for select researchers to edit the files and datasets.

4: Social Science Cases

4.1: Local Elections Surveys (University of Southern Denmark, The Danish National Archives)

Table 4.1: The data involved in the project:

Project:	**Local Elections Surveys**
Type of data:	1 SPSS file
Amount of data	approximately 24 MB

4.1.1: Description of the case

A research group at the Department of Political Science, SDU, has conducted comparable voter surveys in connection with every Danish local election since 1993. Data materials (digital data and documentation) from the surveys in 1978, 1981, 1983, 1993, 2005 and 2009 are preserved long-term in the Danish Data Archives, now Danish National Archives. The case uses the 2013 survey and qualifies the upcoming 2017 survey. The surveys already archived in The Danish National Archives are preserved in DDA's archiving format, DDI Lifecycle, and may be converted to any (major) system format (upon request).

4.1.2: Progress during the project

The case team[44] has focused on:

A. Identifying exact data material for inclusion as case
B. Identifying researcher-perceived benefits and concerns arising from enhanced and present data management practice
C. Identifying deliverables from the project
D. Researcher feedback to DMPonline

Re. a), the election survey from 2013 has been submitted to the National Archives and archived as part of the project.
Re. b), the researchers are very supportive of steps taken in the project towards enhanced data management. They explicitly state that they are in need of a more systematic approach with regard to planning, managing and preserving data collections.

[44] Responsible: Lone Bredahl Jensen, University Library of Southern Denmark (SDUB) and Anne Sofie Fink, National Archives (DDA). Participants: Christian Lindgaard Olesen, National Archives (DDA) and Christina Guldfeldt Madsen, National Archives (DDA) + Ph.D. students and researchers (key contact: Christian Elmelund-Præstekær, SDU).

4.1.3: Questions raised include:

- How to store data while it is actively used in the research?
- When to archive the data?
- Which version of the data to archive?
- How to ensure proper management of personal data?
- How to document data in a series to ensure and to clarify comparability?
- How to be proactive on funder requirements (H2020, Danish national public funders, etc.)?
- How to use metadata to qualify subsequent data collections?
- How to facilitate data discovery for new users?

Re. c), deliverables from the case are:

- Section on data management ('best practice') for inclusion in a list of data management recommendations
- Data Management Plan for election survey 2013
- Initial planning of the 2017 election survey
- Identification of relevant software for managing both research process and data dissemination

The list of data management recommendations has been developed and evaluated by two Ph.D. students at the University of Southern Denmark. In future, the list will be improved and made open to the public.

The DMP has been created in DMPonline[45]. Firstly, DMPonline was applied to document metadata from the finished and archived 2013 survey. Secondly, as part of the initial planning of the 2017 election survey, DMPonline was then used as a planning tool for the future 2017 survey, taking the 2013 DMP as its starting point.

Re. d), we have demonstrated DMPonline to the research team. They are generally positive about the prospects of using DMPonline to create and manage DMPs, but they noted that the standard template may not suit all research projects. Specifically, they would like more predefined reply options to ease the creation of DMPs, mostly for smaller research projects. They would also like DMPonline to be integrated with other services for researchers, such as PURE and the institute's own website, so researchers do not have to provide the same information multiple times. Likewise, incorporating The Danish National Archive's policies for archiving data in the DMP would ease the information burden on researchers. Integration with other services would also make it easier for researchers to make their research projects and data more visible, as well as sharable with other researchers and students.

[45] See Appendix 8

4.2: CALPIU - Centre for Cultural and Linguistic Practices in the International University (Roskilde University Library)

Table 4.2: The data involved in the project:

Project:	**CALPIU**
Type of data:	Video files: MOV Audio file: WAV Total 360 hours of recording. Of these, approximately 28 hours have been transcribed using linking software (CLAN)
Amount of data stored on the project's server	approximately 4 TB
Amount of data which should be stored in a repository	approximately 100GB

4.2.1: Description of the case

CALPIU Research Centre[46] aims to provide an organizational framework for Danish, Nordic and international cooperation for the purpose of creating a new theoretical understanding of the internationalization process which universities are currently undergoing. The centre's main research focus is the function of language in the social and cultural practices of the university, especially the significance of language choice and language proficiency within a context of power relations and hierarchies of influence, as well as the significance of power relations and hierarchies of influence vis-à-vis the organization, didactics, learning processes and academic content of educational programmes in the humanities and social and natural sciences. In all industrialized countries, the language situation at universities is complex and our knowledge of this multilingual reality is rather limited, both in Denmark and worldwide. We know very little about the influence of this diversity on teaching and learning processes.

CALPIU's Storehouse: This project has gathered and processed audio and video recordings of many different types of university activities over a period of three years. These include classroom teaching and lecturing, project supervision and student group meeting activities, student counselling and administrative activities. University ceremonial occasions and speeches are also included in the Storehouse. The intention of the Storehouse project was to gather recordings of naturally-occurring activities and interactions on behalf of the interested researchers and transcribe them in a first-pass procedure which included a basic CLAN transcription prior to further analysis.

Researchers from CALPIU have been involved with the Data Management in Practice (DMiP) project right from the start. CALPIU came up as a potential case when investigating Research Data Management at Danish universities for an earlier DEFF[47] project in 2011-2012. One of

[46] CALPIU continues to exist as a research centre but was mainly active during the period 2007-2011, when it was funded by a grant from The Danish Council for Independent Research, Culture and Communication (Forskningsrådet for Kultur og Kommunikation).

[47] The aim of the project "Management of Research Data (Forvaltning af forskningsdata i Danmark)" as referred in the text, was to identify practices and policies at Danish universities and research funds with regard to the management of research data. Thestrup et al, 2012.

the main motivations for CALPIU's researchers to participate in the DMiP project was their interest in the long-term preservation of their data[48].

4.2.2: Data Management Planning

Roskilde University Library invited the responsible researchers from CALPIU to draw up a Data Management Plan (DMP) using DMPonline. This was done at the end of January 2016. The meeting was also used to obtain feedback from the researchers in terms of the functionality and usability of the tool.

Researchers considered DMPonline to be a good tool to create a Data Management Plan[49]. Overall, the questions were considered relevant, although there was one question that was open to interpretation depending on the background of the researcher, and one that seemed to be politically motivated (OA). The researchers suggested that DMPonline could be improved by asking precise questions, by making some questions obligatory and by having some predefined categories where possible.

With regard to producing a Data Management Plan, it is Roskilde University Library's experience that it is an ongoing process. Although the project period has ended, and the DMP was produced retrospectively, Roskilde University Library still needed to ask follow-up questions to the researchers several months afterwards with regard to aspects such as software versions and system requirements, access control and the organization of datasets.

4.2.3: Data storage, capture and documentation

CALPIU data is stored on a NAS-server, purchased as part of the project. Data setup is located and administered by IT services at Roskilde University. IT services are also responsible for backup and recovery. Data consists of different audio and video recordings, as well as transcription files. Data is organized in a Storehouse and structured according to a 'directory' structure with folders for each subproject and systematic filenames.

4.2.4: Select and deposit for long-term preservation

As CALPIU's researchers are interested in the long-term preservation of their data, Roskilde University Library and the Royal Library arranged a meeting with The Danish National Archives to discuss possibilities for long-term preservation of CALPIU data and the Royal Library's data in The Danish National Archives.

One challenge for long-term preservation of CALPIU data is to manage sensitive research data in such a way that it gives other researchers the possibility to make use of the data in the future. This requires either anonymization of data, or specific preconditions to manage access to the data. If audio and video files need to be anonymized, this will require

[48] Responsible: Janus Mortensen (Associate Professor, Roskilde University/Copenhagen University), Sacha Zurcher and Stine Vejlebo Hansen (Roskilde University Library)
[49] See Appendix 9

resources. These resources are not included in the original project, and there are no special resources allocated to this task. Security and access control with respect to sensitive personal data is therefore essential for the CALPIU researchers in order to be interested in a possible solution that The Danish National Archives can offer.

Another challenge is whether the file formats used by CALPIU can be supported by The Danish National Archives. In this case, it is essential that recordings (audio or video recordings) are linked to transcripts of these recordings. If The Danish National Archives cannot support these file formats, the whole idea of linking transcriptions to recordings will be lost.

CALPIU used CLAN and Talk Bank software. CLAN is one of several programs developed as part of the larger Talk Bank[50] project. CLAN is updated regularly and CALPIU is using the latest version (https://tla.mpi.nl/tools2/tooltype/annotation/computerized-language-analysis-clan/[51]). All versions of CLAN are in principle "downwards compatible". CLAN files can be converted to XML format for future security and "interoperability" via the program Chatter[52], but CALPIU does not have the resources necessary for implementation of this. CLAN uses QuickTime (backend) to play audio and video.

Unfortunately, the type of file formats that The Danish National Archives supports would not make it possible to preserve CALPIU data in its current form, which meant that CALPIU was not able to deposit its data for long-term preservation. Roskilde University Library looked into whether it was possible to use the State and University Library's (now The Royal Library in Aarhus) DSpace Research Data Repository facility, but this was not possible either.

4.2.5: Data identification, citation and discovery

As Roskilde University Library was not able to find a suitable repository for the data of CALPIU, together with the fact that it would require considerable time and effort to anonymize the data, a dataset to which a Digital Object Identifier (DOI) could be linked has not been prepared.

4.2.6: Sustainability

In relation to sustainability of the CALPIU data, it will depend on future possibilities to find a suitable repository that can comply with the criteria of the CALPIU researchers in relation to software used, security and access control.

[50] https://talkbank.org/, accessed 11032017
[51] Accessed 11062017
[52] https://talkbank.org/software/chatter.html, accessed 11032017

5: Health Science Cases

5.1: The GALAXY project (University of Southern Denmark)

Table 5.1: The data involved in the project:

Project:	**GALAXY**
Type of data:	(1) clinical study logs (2) patient history (3) clinical investigations (4) questionnaires (5) outcome event data (6) existing data from the Danish unique personal ID registries (7) biobank logs (8) data generated from quantitative analyses of human liver tissue, blood, faeces, urine, saliva, hair and sigmoid tissue (9) rodent logs (10) data generated from quantitative analyses of rodent liver and colonic tissue, blood and faeces
Data Volume:	We estimate the full dataset to be 1 TB, for which participating centres have sufficient storage capacity.
Number of files:	No estimate

5.1.1: Description of the case

"The vision of GALAXY is to optimize personalized healthcare by stratifying individuals who would benefit from targeted healthcare efforts versus those who would not, and thus to reduce the economic burden for healthcare systems and to improve health outcome for the patients" (http://www.livergalaxy.eu/[53]). The specific targets for the GALAXY project are patients suffering from cirrhosis of the liver, which is most often – but not always – linked to alcoholism. Cirrhosis is considered irreversible, but its precursor, liver fibrosis, is reversible when detected before disease progression. It is the goal of GALAXY to test if improving the gut microbial ecosystem could lead to better interventions and identify new biomarkers for diagnosis, hopefully leading to prevention of progression of the disease.

[53] http://www.livergalaxy.eu/, accessed 11032017

5.1.2: Data Management Planning

During the project, SDUB, The National Archives and Odense Patient Explorative Network (OPEN) worked with the responsible researcher from Odense University Hospital on a Data Management Plan using the Danish installation of DMPonline[54]. The DMP is available at the project website. The researcher considered the tool and the template very helpful and meaningful with regard to the task and valued the support from the National Archives and the University Library in understanding some of the questions (we used the Horizon 2020-template since the project is funded from this programme). This led to a discussion about the guideline texts available through the tool and how these could be improved. One of the things that could be improved was more examples of what an answer might look like in order to clarify what is meant by a particular question. Links to external resources (e.g. Digital Curation Centre (DCC)[55]) would also be helpful. This is an issue that will be considered in future work with the implementation of DMPonline locally at SDU.

Not all questions could be answered right away because the project is at an early stage, but work on the DMP[56] will continue as the project progresses, and we have all expressed willingness to provide further assistance in this regard.

5.1.3: Data storage, capture and documentation

The specifics in this paragraph are only valid for the clinical part of the GALAXY study, which is performed at Odense University Hospital.

Data collection in the project will consist mainly of clinical data from patients undergoing treatment at OUH, but data will also be collected in the form of biological material: Tissue, blood, faeces, urine, saliva, etc. Registration of the biological material will be entered into the OPEN Project's database hosted by OPEN. Data and metadata will be handled using REDCap, which is open source software developed for clinical research by Vanderbilt University. OPEN is responsible for REDCap installation in the Region of Southern Denmark. This system is able to handle metadata, randomize patients, clinical data storage, surveys and more. The REDCap system is designed to handle sensitive patient-level data, since it is encrypted and performs logging of all data entries. Other datasets will be stored in a SurveyXact database hosted by Rambøll and licensed to the Region of Southern Denmark, in secure SharePoint drives hosted by the Region of Southern Denmark and the COSMIC electronic patient file system hosted by the Region of Southern Denmark.

5.1.4: Select and deposit for long-term preservation

Using the REDCap system also enables the researcher to transfer selected data to the National Archives in a structured way. The data has been reported to the National Archives and it is expected that transfer of data can be carried out during the project period, and, in any circumstances, at the latest before the approval from the Danish Data Protection

[54] https://dmponline.deic.dk/, accessed 11232017
[55] http://www.dcc.ac.uk/ accessed 01092018
[56] See Appendices 10 and 11.

Agency expires in 2030. This transfer of sensitive data will also satisfy the demand by law to delete sensitive data after the approval expires.

5.1.5: Data identification, citation and discovery

The goal is to upload anonymized data to Zenodo or another general purpose repository in due time. However, it is difficult to say exactly which datasets and when, since the project is at an early stage. The uploading of data to a repository will most likely take place in connection with preparing manuscripts for publications.

5.1.6: Sustainability

This case has proved to be an important factor in establishing and consolidating the cooperation between the participating institutions – The Danish National Archives, OPEN and SDUB. The case has also given important input to the local SDU Data Management Forum, which is operating on a case-by-case basis until a more permanent support structure can be established. The experience gained here can be fed into the work going on at SDU DM Forum on the development of a data policy at SDU.

Work with the researcher has highlighted issues that need attention both in the project on DMPonline implementation and locally at SDU. These new networks between the project partners are being consolidated and the goal is to streamline them even further. OPEN and SDUB/SDU are working together in the DM Forum, whilst The Danish National Archives and OPEN are collaborating on development of a system of how to report and, if decided, how to transfer data to the Danish National Archives.

This project has contributed to supporting the development of the area of data management, which is set to continue in the future.

6: Science Cases

6.1: CAFF terrestrial CBMP - Technical support to operationalize parts of the Conservation of Arctic Flora and Fauna (CAFF) terrestrial Circumpolar Biodiversity Monitoring Programme (CBMP) (Roskilde University Library)

Table 6.1: The data involved in the project

Project:	**CAFF CBMP**
Type of data:	MySQL Databases: Reproduction, Phenology, Territory occupancy, Banding, Locations, Eggshell Thickness, Environmental chemistry, Geolocator, Prey density. In all 800-900 KB. Files in .csv format.
Raw video material	. mp4. Approximately 20 GB.
Video on YouTube	2 GB.

6.1.1: Description of the case

The terrestrial group under CAFF's Circumpolar Biodiversity Monitoring Program (CBMP) has included the Gyrfalcon and the Peregrine Falcon among the handful of top predators in its plan as "Focal Ecosystem Components" (FECs). The Falcons have, as top predators, a central role in relation to describing the general effects in the ecosystem caused by, among other things, climate change. Falcon species are included due to their well-known role as a monitoring organism, both in terms of the accumulation of pollutants and the effects of climate change in the Arctic, but also because valuable data exists in the form of extended series (> 30 years) with basic population and reproductive data from large parts of the circumpolar area.

This project (period 2016-2018) provides an overview of existing data covering a period of more than 30 years, and will contribute to the planned *State of the Arctic Biodiversity Report* in 2019. It will make 35 years of monitoring data from South Greenland available in a publicly accessible repository and it will try to establish a network of research teams in the circumpolar CAFF area for sharing of the primary monitoring data in the future. This work is a direct contribution to CBMP's work and will be carried out in close cooperation with the terrestrial CBMP group and CBMP Co-chair throughout the process.

The terrestrial CBMP joined the Data Management in Practice (DMiP) project in May 2016 to comply with part of their project goal to "make 35 years of monitoring data from South Greenland available in a publicly accessible repository"[57].

6.1.2: Data Management Planning

A Data Management Plan (DMP)[58] using DMPonline was drawn up in order to get an overview, among other things, of the data collected, formats used and possibilities of placing the data in a repository. The DMP was prepared in consultation with the responsible researcher from the project. The questions were considered relevant. The CAFF terrestrial CBMP project has a variety of data; all data except location data (sensitive data to protect the peregrine falcons) will be made publicly available (OA).

6.1.3: Data storage, capture and documentation

The project has established a database and website at www.vandrefalk.dk[59] for its data and publications. The database is hosted by www.one.com[60]. One.com provides professional backup and data recovery. A selection of the data is OA available either directly at the website, or through small video recordings on YouTube. As the research is ongoing, new data and publications are supplied on a continuous basis. Location data, currently protected by username/passwords, is issued by the owners of the project (Søren Møller and Knud Falk). In future, location data will also be considered being made available to relevant departments of Naalakkersuisut (Government of Greenland) and Danish organisations. There is no need for additional services in relation to the management of this phase of the research data lifecycle.

6.1.4: Select and deposit for long-term preservation

The intention is to deposit data from the project at the Arctic Biodiversity Data Service (ABDS - www.abds.is[61]), the data-management framework for CAFF, the biodiversity working group of the Arctic Council, and its programmes and activities, including the CBMP. It is an online, interoperable data management system that serves as a focal point and common platform for all CAFF programs and projects, as well as a dynamic source for up-to-date circumpolar Arctic biodiversity information and emerging trends. CAFF is based in Iceland.

[57] Responsible: Søren Møller (Associate Professor - Roskilde University Library) and Sacha Zurcher (Roskilde University Library)
[58] See Appendix 12.
[59] Accessed 01092018.
[60] Accessed 01092018.
[61] Accessed 01092018.

6.1.5: Data identification, citation and discovery

Both Søren Møller and Sacha Zurcher attended a DataCite workshop on November 25, 2016, organised by DTU. The aim of the workshop was to add metadata to a dataset and create a DOI. DataCite Metadata Schema 4.0 (http://schema.datacite.org[62]) was used to add metadata to the dataset. A DOI of a sample of the data (video recording of a nest site visit) has been created using the facility on the reference tool Mendeley, which allows researchers to upload raw data from their research, making the research citable. Metadata was added using DataCite Metadata Schema 4.0. Awaiting the possibility to deposit data at the Arctic Biodiversity Data Service, the DOI has been reserved, but not yet been made public.

6.1.6: Sustainability

Since 1981, Søren Møller and Knud Falk have been collecting data on peregrine falcons in South Greenland. The study has focused on population density, territory occupancy, production of young, prey selection, nest-site selection, monitoring of pesticide contamination and reduction in eggshell thickness. Since 1985, breeding peregrines (mainly females) have been banded to collect data on turnover in the breeding population. The current project, aimed at making 35 years of monitoring data from South Greenland available in the Arctic Biodiversity Data Service, will give the collected data a sustainable future.

[62] Accessed 01092018

7: Technological Sciences Cases

7.1: DTU Wind Energy: Meteorological data (Technical University of Denmark)

Table 7.1: Data of the DTU Wind case

Dataset	Data Volume	File formats	Number of files
Type I	1.60 MB	ASCII text, Observed data in TAB file format, Modelled data in LIB file format	402 files
Type II	208 MB	ASCII text, IOC magnetic tape format (GF-3)	456 files
Type III	92 MB	ASCII text in format ESRI ArcGIS ASC grids	8 ZIP files

7.1.1: Description of the case

The meteorology group at DTU Wind Energy has a very large collection of data sets as a result of more than 20 years of research. The data has been captured from different wind-monitoring stations around the globe for many years, making it impossible to repeat or reproduce the data collections. Wind and meteorological data still possesses great value for researchers, industry and government agencies working in the wind power area, for example.

The whole “data bank” can be divided into several types of datasets:

- Datasets released with a publication (e.g. books and reports).
- Datasets published independently, but described in a publication.
- Publications in other formats, e.g. short materials, visuals, software source code.
- Other data sets: Measurement data, experimental data, model results, own standard datasets from models, etc.

There are several challenges associated with this project:

- The data collection includes different types of data, e.g. measurement data, computational models, experimental data, etc., which require different storage, archiving and backup solutions.
- Data is stored in a variety of storage formats: PCs, USB-disks, CD-ROM and Floppy disks, etc.
- There is a lack of standard metadata in this discipline. The metadata associated with this data is based on the vast experience of the Principal Investigator (PI).

7.1.2: Goals

The meteorological sections of DTU Wind Energy would like to document, catalogue and archive datasets to make them openly available wherever possible. The final goal is to preserve this valuable data, but also to expose and make their work more visible.

The main questions of this case project are:

- Which repository/archive would be appropriate for this data?
- Which is the appropriate metadata for these datasets?
- What are the requirements for archiving this data?
- Is it possible to create a metadata catalogue?

7.1.3: Approach

Our office (DTU Bibliometrics and Data Management - BDM) has had three meetings with the PI of the meteorology group which has collected most of the data. Additionally, a regular exchange of information and discussion took place via email.

Meeting 1: the PI gave us a general description of the available data and explained his wishes for the outcome of this case project.

Meeting 2: our office presented the PI with the DTU template for Data Management Plans (DMPs)[63]. We went through the DMP and provided further explanation of the questions where necessary.

Drafting of the DMP was intended to specify:

1. Detailed description of the available data sets.
2. Identify which datasets are eligible for open access and which ones have any copyright restriction.
3. Determine which formats the data sets are available in and if those formats are appropriate for archiving.
4. Identify all the storage sources of the data and determine if all data sets are extractable from there.
5. Identify whether the metadata associated with the data is appropriate and sufficient to archive the data and make it discoverable; this includes exploring whether there are standards available.
6. Identify an appropriate repository/archive solution for this data.
7. Determine what type of licence is required for the publication of these datasets.

The PI filled out the template with support as required from BDM, which also reviewed the plan and added some comments to it.

BDM suggested uploading the datasets to a general repository like Zenodo, wherever possible, in order to test if the features offered by a general repository are relevant and useful when submitting meteorological data.

[63] The first version of the DTU template can be found on the Data Management in Practice wiki: https://sbprojects.statsbiblioteket.dk/display/DAT/DTU_template, accessed 11032017

Additionally, the usability of the datasets submitted to the repository was evaluated by a researcher from a related discipline, but from another department at DTU.

Meeting 3: the main results of the pilot were revised together with the PI, and some general conclusions were discussed.

7.1.4: Expected outcomes

This case project aimed to contribute to the research data management themes by:

- Evaluating the DTU DMP template and giving input to extended written guidance
- Serving as an example of best practice in research data management
- Defining requirements for an institutional repository or catalogue
- Establishing criteria for selecting data of high value for long-term preservation

7.1.5: Results and conclusions

With respect to the success criteria for the different themes in the project, the main results and conclusions are:

7.1.5.1: Data Management Planning:

Considering the vast collection of datasets available at the meteorological section, the PI decided to create a separate DMP for each type of dataset. The work in this pilot was focused on three types of datasets:

1. Type 1: Dataset part of a publication – *European Wind Atlas*
2. Type 2: Experimental data not part of a publication – *Øresund Experiment Data Bank*
3. Type 3: Modelled data not part of a publication – *Wind Resource Map for the WASA Domain*[64]

The respective DMPs[65] were drafted using the DMPonline tool hosted by DTU and using the DTU template and guidelines prepared by our office. Drafting a DMP proved to be useful to:

- Describe in detail each dataset regarding, e.g. formats, file structure and metadata available
- Identify the necessary documentation for the re-usability of the dataset
- Determine under which conditions the data can be shared (licences)
- Identify ownership/stewardship issues
- Prepare the dataset for submission to a repository

[64] See Appendices 13, 14 and 15.

[65] You can read the DMP online: https://sbprojects.statsbiblioteket.dk/display/DAT/DMP+Pilot+Porject+DTU+Wind, accessed 11032017

DMPonline and the template proved to be a valuable tool for the researcher and also for the support services team (BDM). The main feedback received from the PI involved in this project can be summarized as follows:

- DMPonline is a useful and easy-to-use tool. Describing metadata and relevant information about the datasets using this electronic tool is a very helpful exercise.
- This tool should be developed further based on user experience. At the moment, it is a bit difficult to have a clear overview of each DMP.
- The process of preparing a DMP takes time and effort. Not every researcher will be willing to spend time on this.
- DTU should prepare a strategy on how to implement the DMPs and provide incentives for researchers to encourage data curation at the university.
- The DTU template questions generally make sense and the guidelines provided are useful. Some questions were more or less relevant depending on the type of data.

BDM has used the feedback collected in this pilot regarding data management planning to:

- Improve the DTU template and guidelines
- Improve our instructions for the use of DMPonline at DTU
- Provide suggestions to the DMPonline developers at the Digital Curation Centre in the UK (DCC)
- Create a strategy to implement data management planning at DTU as part of the implementation of the "DTU Policy for retention of primary material and data"[66]

7.1.5.2: Data storage, capture and documentation

No activities related to this theme were part of this pilot project because the datasets used are historical data already collected years ago. However, when testing the re-usability of the datasets submitted to the repository Zenodo[67], it was possible to identify the relevant documentation that needs to be provided when publishing this type of data. See the recommendations appearing in section 7.1.5.3.

7.1.5.3: Select and deposit for long-term preservation

The PI submitted datasets Type 1 and 2 to the general open repository Zenodo. The corresponding DOIs are as follows:

- Type 1: *European Wind Atlas*: http://doi.org/10.5281/zenodo.160136[68]
- Type 2: *Øresund Experiment Data Bank*: http://doi.org/10.5281/zenodo.161966[69]

[66] The policy can be found at: https://www.inside.dtu.dk/Medarbejder/Forskning-innovation-og-raadgivning/Forskningsdata/Planlaegning-af-forskningsdatamanagement/Politikker, requires password, accessed 11032017

[67] https://zenodo.org/, accessed 11032017

[68] Accessed 11032017

[69] Accessed 11032017

Zenodo was suggested by our office (BDM), mainly for the following reasons:

- It is a free of charge, open and a general repository
- Data is stored in Europe (CERN Switzerland)
- It is supported by the European Commission and it is the recommended repository for projects which are part of the Open Research Data Pilot in H2020.
- It is one of the alternatives under evaluation as a possible institutional repository for DTU.
- It offers the possibility of exporting the metadata via OAI-PMH, allowing harvesting and transfer of metadata to other data catalogues, for example, once the DTU data catalogue is in place.
- We assume that it is sustainable.

Submitting these datasets to a repository had two objectives:

- To collect feedback regarding submission of a dataset to Zenodo. This would help us to determine if the features offered in the repository are relevant and useful for datasets produced at DTU Wind. In addition, to gather information on the ease of use of the system by researchers.
- To collect feedback regarding the re-usability of the submitted data. We wanted to test if the submitted datasets can be re-used by other researchers from other disciplines.

The PI experience of uploading the datasets to Zenodo was generally positive. His feedback could be summarized as follows:

- The repository was very easy to use. It took approximately 2 hours to upload each dataset.
- The repository provides all necessary options/fields to add relevant metadata and documentation.
- The possibility of selecting a data licence is very important for appropriate citation.
- Drafting a DMP in advance proved to be very useful in order to have all the necessary information describing the datasets at hand.
- The repository is missing statistics about views and/or downloads, which would be a relevant feature for the research group. To our knowledge, this is under development.

The re-usability of the dataset submitted to the repository was evaluated by a researcher from DTU Electrical Engineering who uses meteorological data for modelling energy distribution systems. The feedback was discussed with the PI of this pilot and the following recommendations were agreed upon for uploading historical meteorological data to a repository (in the future):

- The abstract describing the data should answer the following questions: What, when, why and how was this data collected?
- Any additional documentation describing the data, describing the structure of the data and/or any methodology used for the collection of the data which has been previously published should be included in the dataset package. Although referencing the original publication is very relevant and necessary, extracting this information and including it within the dataset will ensure that the dataset is self-contained and re-usable independently of the availability of the original publication.

- When using an encoding format other than commonly used UTF-8 or UTF-16, it should be indicated in the description of the dataset or as part of the provided documentation.
- For easier ingest of data into the analytical software, it is recommended separating metadata from the data itself. For example, in the case of the European Wind Atlas dataset, the file Denmark/ROENNE.TAB contains strings (name of station), tab-separated decimal point values, tab-separated integer variables and comma-and-tab separated integers, which makes it more difficult to start using the file directly.
- When historical data has been collected in a format which is no longer used (e.g. GF-3), the dataset package should include a version of the data in the original format and a version with a commonly accepted standard format (e.g. CSV). The former will ensure research transparency, while the latter will ensure re-usability.
- If the dataset package consists of different directories, a description of the content of each of them should be added in the abstract where the dataset is described.
- Any file naming convention used should be added as part of the documentation.

The dataset Type 3 was not submitted to any repository until the end of this pilot due to the uncertainties regarding the data ownership/stewardship detected when filling in the DMP. The data was collected in South Africa within the framework of a multi partner project coordinated by the Energy Department of South Africa. DTU Wind Energy is one of the project partners. The raw data is currently freely available for download at http://wasadata.csir.co.za/wasa1/WASAData[70] upon user registration. The meteorology group at DTU Wind Energy has used this data to model wind speed, wind power density, terrain elevation and ruggedness index. It remains to be determined whether it is possible to publish these models in an open repository like Zenodo without breaching any consortium agreement. The PI will bring this issue to the next project partner meeting in summer 2017.

BDM will use this feedback to establish general guidelines for uploading datasets to a repository like Zenodo to ensure high quality and maximum reusability.

7.1.5.4: Data identification, citation and discovery

Zenodo has the following advantages that are relevant for the researchers participating in this pilot:

- It provides each dataset automatically with a unique persistent identifier, which allows proper citation.
- It is possible to export the metadata via OAI-PMH, thus allowing harvesting and transfer of metadata to other data catalogues, e.g. to an institutional data catalogue.
- The PI participating in the pilot has an ORCID. However, Zenodo is not yet integrated with the use of ORCID. To our knowledge, this is in development. Zenodo uses the DataCite metadata schema, which proved to be suitable.

[70] Accessed 11032017

7.1.5.5: Training and marketing toolkits

BDM used the experience and feedback collected in this pilot to:

- Improve DTU guidelines and support services
- Provide input to the working group for evaluation of systems for an institutional repository.
- Gain knowledge and create guidelines for choosing data for publishing and archiving

7.1.5.6: Sustainability

Due to the great value of this historical data, it should be archived as long as the original publications are available, i.e. permanently. Therefore the retention time of 20 years offered by Zenodo would not be sufficient. In the event that Zenodo terminates its services, all data will be transferred back to DTU. According to DTU's Policy on the retention of primary materials and data[3], DTU should provide a data catalogue and (long-term) storage for research data. The sustainability plan of an institutional repository at DTU needs to consider that some datasets need to be archived permanently. A business model plan will be created once an institutional repository is established at DTU.

7.2: DTU Space: Geomagnetic data (Technical University of Denmark)

Table 7.2: The data involved in the project

Project:	**DTU Space: Geomagnetic Data**
Type of data:	Real-time data from observatories, measurements of the strength of the Earth's magnetic field, 1 Hz resolution. Data Formats: ASCII and CDF
Data Volume:	1 MB/file, several GB in total
Number of files:	One file per day and station (around 17 stations, some dating back to 1980)

7.2.1: Description of the case

The Geomagnetism group at DTU Space monitors the Earth's magnetic field with a network of 17 ground stations in Greenland, Denmark and the South Atlantic. The main goal in this case is to make the data derived from these measurements available to the scientific community and the public.

The ground stations are maintained by DTU Space with the help of local people who calibrate the instruments once a week. Three of the ground stations are observatories that also include housing facilities. The measurements are made automatically, usually using computer controlled instrumentation. Every few minutes, data is transferred to an ftp

server located at DTU Space, where quality control and different levels of post-processing are undertaken. The data collection and analysis procedures are well established and described in detail in previously published scientific articles[71].

Data from some of the ground stations is already being provided to different repositories from international consortia like INTERMAGNET[72], SuperMAG[73]and the Tromsø Geophysical Observatory[74], as well as to the European Space Agency, ESA. However, these different organizations all use their own calibration methods, quality control measures and analytical procedures, as well as different formats and metadata standards. DTU Space would therefore like to create their own repository, where they can offer "best quality" data from all 17 magnetic ground stations in a complete and consistent way. The data is considered to be highly valuable and of great interest to other researchers and the industry, and possibly also to the general public.

The planned repository should be accompanied by an appealing web interface, where users can browse and search the data. In addition, new visualization tools should be developed and embedded to present selected data, e.g. as an updated map of the earth's magnetic field. The data should be freely available for scientific and private use, whereas commercial use should only be granted on approval.

7.2.2: Goals

The main goals of this case are:

- To increase the visibility of the group's research
- To provide easier access to the "best quality" data
- To make the data citable in an easy and adequate way

The group would also like to track usage of their data. Whether this should be done by requesting users to register for access or by using e.g. download analytics has not yet been decided.

7.2.3: Approach

DTU Bibliometrics and Data Management (BDM) has arranged several meetings with the participating researchers. At the first meeting, the group was introduced to the concept of research data management, and general problems and wishes regarding the visibility and accessibility of the data were discussed. Based on this discussion (and many other interviews that BDM conducted with researchers throughout DTU), BDM developed an institutional template for Data Management Plans (DMPs) and implemented it in the DMPonline tool. The template was presented to the group at a second meeting and the relevant questions were discussed in detail. The group filled out the template themselves

[71] More information can be found on the group's homepage: http://www.space.dtu.dk/english/Research/Scientific_data_and_models/Magnetic_Ground_Stations, accessed 11032017

[72] International Real-time Magnetic Observatory Network http://www.intermagnet.org, accessed 11032017

[73] Observations of the global magnetic field http://supermag.jhuapl.edu/ accessed 11032017

[74] UiT – The Arctic University of Norway http://flux.phys.uit.no/geomag.html, accessed 11032017

and BDM revised the plan afterwards[75]. The final DMP was used to specify the requirements in more detail and to create a comprehensive roadmap for the case.

The considered amounts of data are relatively small and can easily be handled by the research group. There are three researchers involved in the case: the group leader and two senior scientists, who are responsible for data management, including uploading, monitoring and curating the data and maintaining the measuring instruments. All data considered here is owned by the department.

7.2.4: Requirements

The required infrastructure consists of three layers:

1. Storage for the raw data and an active space for handling the data (different levels of quality control and analysis). Existing servers at the department are used and will continue to be used. They provide automatic backups. Currently, quality control and analytical procedures are performed manually. In the future, automatic workflows are due to be integrated, but there resources are not available at present.
2. A data management system that provides structured metadata and registers DataCite DOIs. This is not yet in place, but different solutions are currently being tested and evaluated at DTU.
3. A new website where users can search for specific time intervals and locations and get access to the "best quality" data. The website should include real-time visualizations of the raw data. Details on this part of the case were presented at the DeIC conference on 4 October 2016[76]. BDM has put the research group in contact with an eligible software engineer and is following development closely. This is ongoing work and is expected to be finished in spring 2017.

7.2.5: Outcomes

The project consolidated BDM's role as the central support function for research data management at the university, in particular for communicating between the researchers and relevant stakeholders in central administration such as the library and IT services, and coordinating all related activities. It has also helped to develop local support functions and training material.

With respect to the success criteria for the different themes in the project, the outcomes are:

[75] The final DMP can be seen in Appendix 16 and here: http://sbprojects.statsbiblioteket.dk/download/attachments/27430917/DMP_updated_2016-07-21.pdf, accessed 11032017

[76] The full presentation is available here: https://www.deic.dk/sites/default/files/uploads/konf-sem/konference-2016/Falco%20Jonas%20H%C3%BCser_DMiP_DTUSpace.pdf, accessed 11032017

7.2.5.1: Data Management Planning

The researchers have used the local installation of DMPonline to write a Data Management Plan. An institutional template and guidelines were created by BDM and implemented in DMPonline. BDM collected feedback on the use of the tool, as well as on the structure and content of the template and guidelines.

7.2.5.2: Data capture, storage and documentation

Existing infrastructure at the department for the storage and management of active data is used. Collection methods and documentation procedures are well-established and largely automated. Once additional resources are available, procedures for quality control and analysis will be integrated. This is completely up to the researchers and does not require help from BDM. However, observations and experience in this case help to define requirements for new IT systems and services for the management of active data at DTU. This is an ongoing activity by a working group consisting of BDM and members from IT and the Library. The geomagnetic data from DTU Space might serve as a test use case for new systems in the future.

7.2.5.3: Data identification, citation and discovery

The possibilities and challenges of citing individual datasets from real-time data have been discussed on several occasions with the researchers and different experts in the area of persistent identifiers. The data also served as a use case in a DataCite workshop held at DTU. A concept for using DataCite Metadata Schema 4.0 and for assigning DOIs to the data has been drafted, but is currently lacking the underlying technical support to be implemented. It is expected that this will be available in summer 2017. All researchers involved in this case have an ORCID.

7.2.5.4: Select and deposit for long-term preservation

All documentation and file formats comply with common standards in the geophysical community, which researchers expect to be the best-suited for potential reuse. Since this case deals with data from an existing and ongoing collection that was established several decades ago and has since been continuously improved, it did not require any further steps to be undertaken in this theme.

DTU will ensure long-term preservation of the data. A final decision of how this will be done in detail has not yet been made. A working group consisting of BDM and members from IT and the Library is currently looking into different solutions for an institutional data catalogue and archive. A recommendation will be presented to the Dean of Research in summer 2017.

7.2.5.5: Training and marketing toolkits

BDM has prepared several guidelines for research data management and shared relevant information with the project[77].

[77] DTU Data Management Plan doi.org/10.6084/m9.figshare.4003857(accessed 11032017) and DTU Research Data Life Cycle doi.org/10.6084/m9.figshare.4258019, accessed 11032017

7.2.5.6: Sustainability

The questions on sustainability are closely related to the plans for long-term preservation of the data and the establishment of a common data catalogue and archive at DTU. This means that final decisions have not yet been made. It is clear, however, that the researchers themselves are interested in maintaining good quality and access to their data and will take responsibility for this, at least as long as measurements are ongoing, also as part of obligations towards partner organizations in the international community. On the other hand, DTU guarantees long-term preservation of the data. BDM is in continuous contact with the institutional IT that will provide the required services and infrastructure.

7.3: The Kepler Case

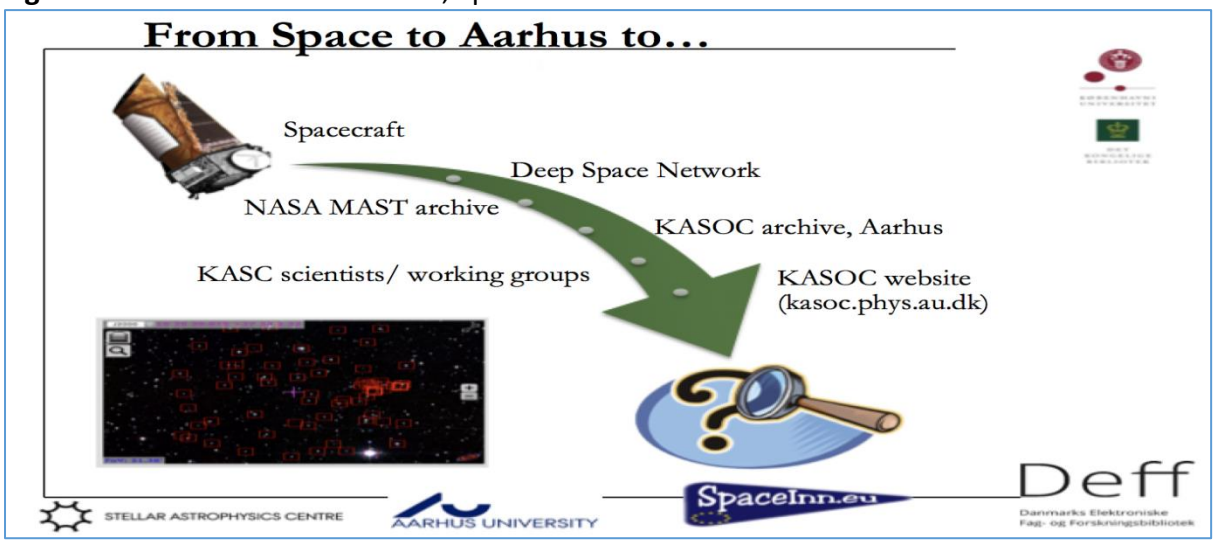
Table 7.3: The data involved in the Kepler project

Project:	**Kepler**
Type of data	One BagIt preservation package for each star, containing all the data files from KASOC for that particular star. This includes up to five levels of data or data products, covering different levels of processing and researcher involvement: 1. Kepler original pixel data (as retrieved directly from the MAST archive) 2. Light curves (time series) originating from NASA as well as from KASOC 3. Power spectra 4. Analyses based on light curves or power spectra (e.g. stellar oscillation frequencies and stellar properties). These analyses are typically published in an article. 5. Physical models of the stars, indicating star mass, size and age, based on observed oscillation frequencies. Inferences drawn from these models are typically published in an article, but the models themselves are typically not.
Amount of data which should be stored in a repository	> 120 TB of Kepler data, but with zip file compression of BagIt packages considerably less TB.

The aim of the NASA Kepler/K2 mission is to observe extrasolar planets orbiting stars outside our solar system. The Kepler spacecraft was launched in 2009 and is still an active photometer telescope floating in gravitational space between the earth and the sun, sampling image data and light emissions in the habitable zone from stars which in terms of size and age have similarities to earth. The management of the +100 TB of data produced in the research use case of Kepler/K2 has been subject to thorough documentation delivered during the period of the Data Management in Practice project [78].

[78] To avoid further repetition we refer the reader to the following final documents providing a detailed description and presentation of the joint work of the Stellar Astrophysics Centre at Aarhus University and the Royal Danish Library: Poster: Conrad, Svendsen and Handberg, 2016; Handberg, Houdek, Christensen-Dalsgaard, Conrad and Svendsen, 2017; Practice paper: Conrad and Svendsen, 2017;
Paper: Conrad, Svendsen and Handberg, 2017.

Figure 7.1: Back UP To The Future, SpaceINN & HELAS8 Conference 2016.07.11-15



However, the Kepler use case[79] is still a work in progress and there are some issues to be resolved ahead of securing a long-term KASOC archive that will be given additional attention in the following.

Each dataset, containing all the available data files from the KASOC database relating to one observed star, will be considered as one data object in the chosen repository and given one identifier (DOI) covering the entire dataset and providing it with citation metadata via the DataCite schema[80]. This will make it slightly harder to cite a single file within a dataset. We have not been able to circumvent this limitation because it would have been a major task to treat each single file as an independent data object, whilst still supplying the necessary context and documentation. The datasets are packaged in the BagIt archive format, which is simple and easy to parse for both computers and humans, and described as being well-suited for digital preservation purposes[81]. BagIt archives can be stored and transmitted as compressed ZIP archives.

The requirement that data must be useful and understandable is central to continued use, and is dependent on data identification and discovery by scientists in the future. This relates to the metadata schemas applied to the datasets and the discovery features being offered by the data repository service. It is anticipated that future discovery of archived Kepler data by researchers, who would not know of the Kepler/K2 mission, will take place by means of the name or more likely by the position of the observed star in the sky. This calls for the use of a discipline-specific metadata standard offered by the IVOA resource metadata specification schema[82], as well as the citation metadata already mentioned.

[79] You can read the DMP of this case in Appendix 17.
[80] Datacite Working Group 2016. See http://doi.org/10.5438/0012, accessed 11062017
[81] Kunze, Littman, Madden, Summers, Boyko and Vargas, 2016, https://tools.ietf.org/pdf/draft-kunze-bagit-14.pdf, accessed 11032017
[82] International Virtual Observatory Alliance: http://www.ivoa.net/xml/index.html, accessed 11032017

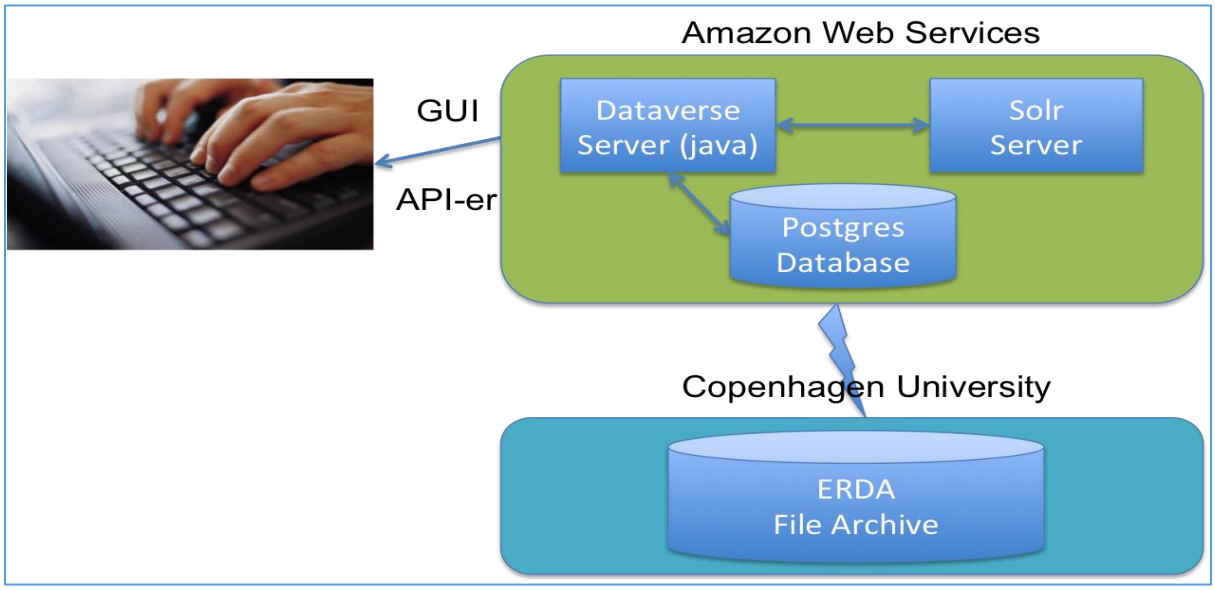
In addition, it raises the requirement for a repository system that allows for discovery based on astronomy-specific metadata allowing for a range search on celestial coordinates in the sky. Although all metadata can be scripted from the KASOC database as a JSON file and ingested in the repository via the API, and metadata can be modelled in a combination of DataCite and IVOA, this single requirement has proved to be one of the most difficult to solve in our work with the case.

7.3.1: Infrastructure perspectives

The Kepler and Kierkegaard cases raised requirements for supporting technical infrastructure for archiving/preservation and access. Due to the fact the Royal Danish Library, Copenhagen, already has a running version of an old version of Dataverse, we wanted to test the newest version of Dataverse as a possible data repository solution. This was made more interesting due to Dataverse's support for astronomical and astrophysical metadata, which was essential to our Kepler case.

The Kepler case will potentially reach about 110 TB of data in total, more than could be stored in freely available repositories or infrastructure that is immediately accessible to us. We therefore established collaboration with the ERDA archive at Niels Bohr Institute, Copenhagen University, which is running a research data archive of currently around 13 Petabytes and would be willing to host the data. We built a test setup of Dataverse for the pilot cases of the project, with two servers and a Postgres database located at Amazon Web Services EC2[83] coupled with the ERDA archive as file storage. We initially connected Dataverse with ERDA through their WebDAV API. We agreed with the ERDA people to consider this as a proof-of-concept installation.

Figure 7.2: Dataverse proof-of-concept architecture



The proof-of-concept setup has worked very reliably with test examples covering our cases. It is slow, however, due to the minimal computer resources we have invested at AWS. In particular, the WebDAV pseudo-file system works with intermediate caches, resulting in a

[83] Amazon Web Services: https://aws.amazon.com/ ,accessed 1103027

number of superfluous write operations when ingesting. In light of the fact that we might eventually ingest many terabytes of Kepler data, this issue has to be addressed. We have agreed with the ERDA administrators that we will be trying other possibilities, eventually arranging for a more tailored connection between our Dataverse servers and ERDA, if necessary.

Dataverse is being run at many sites around the world, and the original developer, Harvard University, has documented its clustered setup. So we assume it would be possible to build a large-scale version of our pilot setup, and we may want to conduct performance testing on a larger setup if time permits.

In order to move to a production setup, we will need a technical partner to run the Dataverse server(s) and associated services, in addition to the storage. Talks are ongoing with Copenhagen University IT Department about them running the Dataverse technical stack, while Copenhagen University Library (part of Royal Danish Library) will retain ownership of the service, as well as the service and repository management functions.

Recently, the Danish Data Archive has shown interest in testing Dataverse support for social science surveys. The proof-of-concept installation has been expanded with extra functionality for this kind of data, which has been tested by Danish Data Archive. Furthermore, another university library has shown interest in testing the platform with social science data from their researchers, allowing us to test how easy it is for an ordinary non-expert researcher to use the graphical user interface.

7.3.2: Conclusion for Royal Danish Library, Copenhagen, cases

In our project cases we have demonstrated that a lot can be done to prepare for long-term stewardship and access for researchers, even though the current institutions and infrastructure are not quite ready for this on a larger scale. So while we have not been able to finish our cases completely during the project period, major steps have been taken, both in structuring and preparing the data, and towards creating the supporting infrastructure.

In our cases we have found that combining the researchers' understanding of the data and scientific domain with data librarian and IT skills results in a very successful mixture of competences for establishing good data management and data stewardship. Leaving out any of these perspectives in a complex case could result in preserved data being either incomplete or useless for research.

The interest from various parties in the Dataverse installation, both on the user and infrastructure side, seems to indicate that it may be possible to solve data management challenges in partnerships and to build infrastructure solutions across institutional boundaries. This will include the need to explore new business models, allowing each partner to do what they are best at. We will keep exploring these perspectives with the interested parties and hope to bring a data repository service that can host all the Kepler and Kierkegaard data into production as soon as possible.

8: Common conclusions of the cases

The different cases described in sections 3 to 7 cover the areas described in the project application: Humanities, Social Sciences, Science and Technology. During the course of the project, a case from Health Science was integrated into the case portfolio. The nine cases cover a wide variety of data types and sizes of datasets. The data varies in size from MB (The Local Elections Survey) to TB (the data analyzed in the Netlabcase). Several cases also demonstrate how legal issues can affect the storage and dissemination of research data and how such issues can be resolved.

The application required that the cases should cover the entire lifecycle of research data. The cases fulfilled this demand. Some cases were part of the start of the research project, such as the GALAXY project. Others did not come into contact with the different projects until late in the cycle, such as the LARM Case or the CALPIU case.

In general, the cases demonstrated that the researchers considered Data Management Planning to be a very useful activity and were satisfied with DMPonline. As required, work on the cases was carried out in close cooperation with the researchers in order to evaluate the existing template in the current DMPonline installation and to develop two additional templates. One was in English as a proposal for a national template and one was to be used specifically in relation to the cultural heritage cluster.

We acknowledge that nine cases cannot claim to fully describe or in any way cover all research areas, but in our opinion they can be regarded as providing sufficient background for further assessments of the problems researchers can meet and services which can be offered to researchers to meet such demands.

The cases showed that the research libraries, the national library and the national archive could all play a vital role in data management. However, the cases also showed that this role will vary according to the institutional framework. In some cases, the libraries and archives collect the research data and research objects, give access to these and operate online repositories. In other cases, the libraries have a more advisory role.

As previously shown, work on the cases resulted in the development of practical solutions and services. A Danish DMPonline server was established, different templates – including a national Danish template – to be used on this server were designed and a model agreement to handle legal issues was completed.

In general, the cases fulfilled the requirements of the project. In the next chapter, we will discuss how the lessons learned from the cases relate to the themes of the project.

9: Thematic areas

9.1: Data Management Planning

Roskilde University Library is responsible for the Data Management Planning theme within the Data Management in Practice (DMiP) project. Based on conclusions from a former DEFF (Denmark's Electronic Research Library) project, Fælles Infrastruktur for Forskningsdata (Common Infrastructure for Research Data)[84], it was decided that DMPonline would be used as a tool to assist researchers in creating Data Management Plans (DMPs). The Digital Curation Centre (DCC) in the UK has developed DMPonline. It helps researchers to create personalized plans according to their context or research funder, and provides examples of guidelines and best practice.

9.1.1. Danish DMPonline installation established and available

Roskilde University Library organised a Master Class on DMPonline in November 2015 for all participating institutions in the DMiP project. The aim of the Master Class was to become familiar with DMPonline in order to be able to assist researchers in using the tool to create DMPs for their research data. Teachers at the Master Class were Sarah Jones and Marta Ribeiro from DCC, both experts who have helped to develop the tool and who are familiar with the system in practice.

There were a total of 22 participants from Aalborg University Library, Copenhagen Business School, Copenhagen University, Technical University of Denmark (DTU), The Danish National Archives, The Royal Library, University Library of Southern Denmark and Roskilde University Library. The participants were all aiming to be super-users of DMPonline to support researchers in preparing DMPs.

Two programmers from DTU with experience in Ruby-on-Rails also attended the Master Class, as DTU had assumed support and operation of the software during the project period.

A Danish pilot of DMPonline was established at a DTU server in November 2015. All those who had participated in the Master Class and those who were part of the DMiP project were registered as administrative users. Researchers could also register and would be regarded as regular users in the system.

9.1.2. DMPonline adapted to Danish researchers and Danish conditions

To be able to assess the functionality and usability of DMPonline, it was necessary that each case from the respective universities prepared a DMP of the research project. Based on experiences from the participating university libraries and researchers, the questions in DMPonline were discussed in a Danish context. Here it became clear that there were differences in the approach of the university libraries and the type of research projects they were dealing with.

[84] Common Infrastructure for Research Data was a DEFF project, which ran from April 2013 to July 2014. See Larsen et al, 2014.

Natural science and technical research projects considered the questions relevant and adequate. Social sciences and humanities also regarded the questions relevant, but needed more specific questions related to legal rights and obligations in relation to sensitive data. In addition, some participating university libraries were not sure whether all researchers would be familiar with the terms used in DMPonline. This could be solved by adding links and guidelines to the questions.

It was agreed that there should be one template for all universities, instead of separate templates for the natural sciences, technical sciences, social sciences and humanities. A request from the participating university libraries to develop a Danish template which was organized in phases was not possible to implement. DMPonline does not automatically transfer the answers of a previous phase to the next phase. This means that in a new phase it would be necessary to repeat the answers given in an earlier phase. As this meant more work for the researchers, the idea of a template in phases was abandoned. It was agreed that there should be a Danish version of a DMP. Each university would be free to use the Danish template, or a template designed specifically for their university. As of this moment, DTU and the Royal Danish Library, Aarhus are the only organizations to have developed their own DMP template.

Roskilde University Library prepared a draft DMP based on Skype discussions with the participating organizations. Many questions in the Danish template were inspired by questions from DCC's DMP template, but made more specific in the Danish version based on requests from the participating university libraries. This resulted in considerably more questions in the Danish version than in the DCC version. The Danish version of a DMP consists of an introduction on Plan details and 6 sections. These are: Data collection; Documentation and metadata; Ethics and legal compliance; Storage, backup and security; Selection, preservation and sharing; and Responsibilities and resources. Each section has a short introduction followed by a number of questions. The Danish version has a total of 60 questions, which for some might make the DMP too extensive.

In the final recommendation to the steering committee of the DMiP project, Roskilde University Library and DTU (the host University for DMPonline), recommended that the Danish template be implemented as a funder template. The advantages are that it is possible to create two templates – one in English and one in Danish – in addition to which each university would be able to include specific guidelines to each question and to customize the template by adding extra questions and sample answers. The disadvantage is that it could create confusion amongst users to find the Danish template under funders. On February 24th the steering committee decided to follow this recommendation. The Danish e-Infrastructure Cooperation will host DMPonline from April 2017.

9.2: PURE and research data management

In addition to the themes and cases, we have looked at how PURE (the institutional repository software provided by Elsevier) can be utilized in the field of data management. The reason for looking specifically at PURE is due to the fact that this system is in place at all participating institutions as the main repository for articles.

To gain insight into the topic, we asked three UK institutions and a representative from Elsevier to each create a webinar for the Danish universities. The following participants each gave a presentation:

- Masud Khokhar, Lancaster University
- Anna Clements & Federica Fina, St. Andrews
- Kerry Miller, Edinburgh University
- Henrik Rasmussen, Elsevier

This – combined with the insights from using PURE as an institutional repository – led to the following points of awareness. Please note that these observations are from the spring of 2016. Some of the issues raised may have been fixed in a later version of PURE; however, the overall conclusions should still be valid.

9.2.1. Registration in PURE

PURE has a nice familiar interface for registering data, and facilitates upload of files. However, it lacks certain functionalities, e.g. recognizing files, and everything is done through a browser. Although PURE has the ability to store research data files, it is not the optimum system for such data, particularly when it comes to data preservation and curation activities. There is also no versioning for datasets. The metadata model is very simple (although some users perceived it as quite complex in terms of the user interface), which might also reduce the usefulness of the metadata that can be stored in PURE.

A dataset can be quite hard to define. In terms of datasets in PURE, the dataset is an abstraction for a larger set of data with a common denominator. Some researchers produce twenty datasets in a single day, some produce one a year. The current practice is to register datasets in PURE related to publications, which yields a couple of hundred datasets a year. Storing metadata for these datasets in PURE can hold great benefits. It allows the end-users (whether these be researchers, editors or administrators) to relate datasets to other PURE content. This can include articles, projects, equipment, etc., which can then be further linked to awards and so on. It is then possible to include datasets in presentation profiles of researchers, projects, etc., in the public portal.

The work processes would be familiar to the users (including administrators) and be able to provide a good platform if a trusted system is to be built. However, the limitations of PURE for actual storage of research data have to be kept in mind. The recommendation is not to keep research data in PURE, but to use PURE for registering metadata and related datasets. There is currently a bridge between Mendeley Data and PURE for pushing metadata from Mendeley Data to PURE.

An issue raised at the webinars was the ability to make sure that APIs could write data to PURE. This is almost impossible in a generic way. So other systems might struggle with providing data to PURE using APIs (e.g. for automated registration in PURE based on actions in a different system). Some import modules might solve this problem, but these are not APIs as such.

PURE has some issues on DOI handling, e.g. the fact that removing a dataset from PURE that has a DOI issued will result in an HTTP 404 – Page not found error, whilst the transfer of metadata from PURE to DataCite has an issue related to the conversion of metadata from PURE to DataCite, causing records to have shortcomings in terms of completeness.

9.2.2. PURE as a Data Source

Depending on the institution's use of PURE for registering projects, awards etc., PURE might be a good source for data evaluation of the content of other systems that deal with data management. If PURE is a source for knowing all current projects within an institution, data from PURE can be used to evaluate content in other systems. An example could be DMPonline. If there is a university policy stipulating that each project must have a Data Management Plan, how could this be verified? If PURE can provide a list of current on-going projects and DMPonline can provide a list of projects with DMPs, the remainder (i.e. projects without a DMP) can be found by comparing these two systems. However, this is only true if the project can be identified in both PURE and DMPonline, e.g. using a shared key to identify the project.

9.2.3. PURE as a repository for documents related to data management

If there are documents related to data management within a project group or similar, e.g. a Data Management Plan, PURE will have the ability to store the document and link it to a dataset, projects, etc. The same lifecycle for showing/hiding content, etc., as any other element in PURE and making workflows for this type of content are also possible.

9.2.4. PURE for registering data management activities

With activity registration in PURE, the system can be used to register activities related to data management and have these activities tracked and/or exposed in the profile of the user. Thus in a case where a person has contributed to the creation of a Data Management Plan, this work can be registered in PURE. This allows PURE to keep track of activities related to data management.

9.3: Data capture, storage and documentation

This stage of the research data lifecycle is understood to be part of the active research project – i.e. how researchers deal with their data during collection and processing, and – very important in terms of long-term stewardship – how data is documented for possible later reuse.

The cases in this project have tended to use very diverse solutions to this lifecycle step, ranging from using existing research infrastructure (e.g. LARM, Local Elections Surveys), through project-provided or university/department-provided infrastructure (e.g. CALPIU, Kepler, DTU Wind) to cloud-based storage (e.g. CAFF terrestrial CBMP). In some of the cases, the data had already been collected before the project case started, making this step of the lifecycle less relevant to deal with.

The project cases have therefore not resulted in any coordinated action in this area, and it would not seem obvious from the project what kind of common infrastructure might be needed, or whether the libraries would play any role in providing it.

9.4: Data identification, citation and discovery

9.4.1. Background and success criteria

In theme 3 we explored how datasets can be identified, cited and discovered. Initially the following assumptions were made:

- Researchers would have ORCID identifiers
- Datasets are expected to be identified by DataCite DOIs or DanPIDs
- Metadata about datasets should be harvested from local repositories as well as national (long-term storage) archives (e.g. the Danish National Archives)
- Datasets will be searchable in the Danish national research database (www.forskningsdatabasen.dk[85]) and in international databases.

These assumptions led to the following success criteria:

- All datasets must have metadata containing a unique identifier (DOI or DanPID)
- All subjects related to a dataset must be represented by a unique identifier, i.e. ORCID
- All datasets must be harvested and indexed in the Danish national research database

Moreover, it was stated that cross-linking between researchers, publications, datasets and other related digital resources (e.g. by visualisations) is explored by modelling examples for each use case in the project.

[85] Accessed 11032017

9.4.2 Revised success criteria

The assumption that institutions would have local data repositories did not reflect reality at most of the partnering institutions. Only the Royal Danish Library in Aarhus managed to implement a data archive based on DSpace prior to the end of the project. In addition, a number of institutions led by Aalborg University Library have investigated using PURE as their data repository. Final recommendations are in preparation.

Another assumption was that the Danish National Research Database would be able to harvest datasets from data archives by the end of the project. However, the Danish National Research Database has undergone a major change in its governance due to a change in ownership (transferred to the Danish Agency for Science and Higher Education) and establishment of an advisory board. Consequently, the inclusion of datasets in the database was not possible to implement within the active project period.

This situation has led to the following recommendations:

1. Datasets must provide metadata based on DataCites Metadata Schema version 4.x
    - This can be done through a data repository that provides DataCite metadata for external harvest. Local repository or external general purpose repository (e.g. Zenodo, FigShare, etc.).
    - This can be done by providing XML based on the DataCite metadata schema, either manually or by using one of the DataCite metadata XML-generator forms[86].

2. Datasets should provide a unique identifier, preferably a DOI
    - If institutions do not have DataCite membership, obtaining a DOI can be achieved by submitting the data to general purpose data repositories like Zenodo or FigShare.

3. All researchers related to the dataset should have an ORCID
    - It is free to obtain an ORCID for researchers – however, it is completely voluntary for them to do so; therefore this can only be highly recommended – not mandatory.
    - DataCite metadata schema allows Creator(s) to be associated with an identifier, including ORCID.

At present, the Danish National Research Database does not accept datasets. The project therefore makes two recommendations. The first recommendation is that datasets are provided with DataCite DOIs, since this also requires the metadata to be submitted to DataCite and therefore available for search and discovery in the DataCite search portal[87] . Danish Universities will soon be able to mint DataCite DOIs at no cost through DeIC increasing accessibility to this type of persistent identifier. Considering that DanPID has been used mostly as a persistent identifier by Danish cultural heritage institutions, it indicates that DataCite DOIs are the best recommendation for research data. Secondly, the

[86] Example of DataCite metadata XML generator: http://dspace.ut.ee/handle/10062/53326, accessed 11032017

[87] DataCite Search: http://search.datacite.org, accessed 11032017

project recommends that the Danish National Research Database adopts the DataCite metadata schema as its exchange format for including datasets in the database.

9.4.3 Linking and visualisations

Using the DataCite metadata schema also allows for the creation of links between different related information objects using the relatedIdentifierType field and semantic expressions like IsCitedBy, IsPartOf, IsDerivedFrom (see page 37 in the DataCite Metadata Schema Documentation for the Publication and Citation of ResearchData. Version4.0[88]). The project therefore recommends that relations between datasets and other related information objects are made using the DataCite metadata schema. This can be done either via metadata alone or through a simple graphical visualisation (e.g. hand-made object model like diagram).

9.4.4 DataCite – recommended as exchange format

For data identification, data citation and discovery, the DataCite metadata schema is well suited.
The project recommends:

- Adopting DataCite metadata schema as the recommended exchange format for exchanging metadata with the Danish National Research Database.
- Not to create a national application profile for DataCite metadata schema
- To nominate a Danish member to the DataCite Metadata Working Group representing Danish interests and experiences.

The mission of DataCite is to make data citable through the provision and minting of unique identifiers (DOIs) and metadata for discovery. Accordingly, the metadata schema is intentionally produced for general purposes and suited to any kind of data repository. The trade-off for this is constraints on the depth and scope of the metadata schema. In particular, the Kepler case from the Royal Library was challenged by the limitations of the DataCite schema that did not include space coordinates for objects in space such as planets. Other constraints could be localized data and data fields only relevant for Danish institutions. In light of these constraints, it was discussed whether it would be useful to have a Danish application profile for the DataCite metadata schema, like the OpenAIRE Guidelines for data archives[89] . Keeping a local application profile for the DataCite metadata schema raises a sustainability issue, however, since an application profile adds complexity to data providers and it would require resources to sustain and support a national application profile. Instead of having a national application profile, the project recommends that the national community works towards and supports nominating a Danish member to the DataCite Metadata Working Group[90] .

[88] https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf, accessed 11062017
[89] OpenAIRE Guidelines for data archives: https://guidelines.openaire.eu/en/latest/data/index.html, accessed 11062017
[90] Info about the DataCite Metadata working group is found on the page: http://schema.datacite.org/, accessed 11062017

In order to obtain practical experience or gain hands-on experience using the DataCite metadata schema, a workshop was organised with Laura Rueda from DataCite at the Technical University of Denmark on November 25th. Members from all institutions participated – in total 15 people attended the workshop.

9.4.5 Results

Since there is still a lack of institutional repositories with full implementation of the DataCite metadata schema and DOI integration, progress in the different cases varies. However, some institutions have either demonstrated proof-of-concept for an institutional repository (DSpace at the Royal Library in Aarhus and DataVerse at the Royal Library in Copenhagen) or are currently investigating different solutions (e.g. FigShare for institutions at DTU) where integration of DataCite and ORCID are essential requirements, as well as the exchange of metadata with the institutional research information systems. There is constant sharing of knowledge and experience between the different institutions.

Several cases have published datasets in general repositories (e.g. Zenodo) or are soon due to publish datasets in discipline-specific repositories using the DataCite metadata schema and DOIs.

It is clear that work related to this theme is ongoing. However, the importance of standard metadata and persistent identifiers for datasets and authors (researchers) has become obvious for everyone involved in the cases. It is expected that recommendations concerning their use will be included in the implementation of research data management policies and education programmes at all universities. Their success, however, will inevitably depend on the availability of supporting infrastructure.

9.5: Select and deposit for long-term preservation

9.5.1: Success Criteria

This phase describes the end of a research project when it is time to start preserving data for long-term access and re-use.

It covers criteria that will help to:

- Select suitable data and contextual material
- Prepare data for preservation
- Evaluate and choose repository
- Recommendation for long-term preservation

This is all documented on the project wiki[91]. The highlights and the work process will also be covered here. Work on the evaluations combined with the needs in the different cases pointed toward establishing a new research data repository pilot service. This was done

[91] https://sbprojects.statsbiblioteket.dk/display/DAT/Data+Management+Home, accessed 01092018

towards the end of the project. The new pilot is an open access repository. Some of the cases further pointed to the need for a restricted access repository service, which we will also discuss here.

9.5.2: Experiences

When working on this theme, it became clear that the different cases posed very different requirements to long-term preservation: Some cases were tailored to using The Danish National Archives for deposit; some cases were expected to use an institutional repository; others were looking for a national open access solution, and this theme set up a pilot open access repository for research data, which will be described below.

A lot of the cases, however, turned out to contain sensitive data, which requires restricted access. The Danish National Archives provide a restricted access repository, but it is not suited to all types of data. We therefore propose a new service for this, also described below.

9.5.3: Success Criteria 1: Documentation and Best Practice

- Develop relevant documentation and best practice
- Develop criteria for how data is to be preserved in the long term
- Assessment: Documentation exists

These success criteria are largely related to the work on the wiki and are documented here.

9.5.3.1: Select suitable data and contextual material

When it is time to preserve your data, you will need to consider carefully exactly which components of your research need to be preserved. This is a complicated process and depends on the research area. There are of course also general guidelines. The most important guidelines are that you should *definitely deposit* original data. It is *not necessary to deposit* data that is already preserved and accessible via other institutions or organizations. And *do not deposit* – or at least be very careful if you do and ensure proper restrictions when depositing – data that contains sensitive personal data or copyrighted material. The last point has been important in many of the cases, as some of the data involving human subjects is important original data which should certainly be preserved, but which certainly should also have restricted access.

9.5.3.2: Prepare data for preservation

Preparing data for preservation and reuse is not a stage, but an ongoing part of the research process. When preparing data for preservation and re-use, remember the following points:

- Archives and repositories require clarity on who owns data and that permission for preservation and re-use is granted. You usually accept this when accepting the distribution licence.
- Preparation of sensitive data involves potentially removing sensitive data, anonymizing it and matching the data to an appropriate licence or level of controlled access in the archive or repository.
- Documentation and Metadata: Data requires good explanatory contextual material and information to be accepted into an archive or repository.

- Converting or migrating data to enable the data to be preserved for the long term may also be necessary. Check the integrity of converted files as thoroughly as possible immediately afterwards.

9.5.3.3: Evaluate and choose repository

From the large repository list at re3data.org[92] , we have examined the following repositories:

- Global repositories[93]
  - Zenodo
  - figshare
  - The Dataverse Project
- Danish repositories[94]
  - Databox (based on Dataverse)
  - Danish National Archives
- Local institutional repositories[95]
  - CKAN
  - DSpace

We established a long list of criteria that we used to evaluate each repository. We note that both the list of available repositories and the list of important criteria change rapidly. Our evaluation can be used as inspiration, but we recommend a new evaluation, if not with every project, at least at fairly short intervals. Also, the decision on which repository is the best depends very much on your community and the data you wish to preserve.

9.5.3.4: Recommendation for long-term preservation

Project website or informal sharing is not recommended for long-term preservation or re-use of data, as lack of maintenance is a significant risk. Funders, institutions and journals may specify where data should be placed in specialist data archives, platforms, or institutional repositories. In general, authors are encouraged to archive data in disciplinary data-type specific repositories, where preservation is most likely the best and popularity in the discipline highest. However, generalist repositories can handle a wide variety of data, and may also be appropriate for storage of associated analyses or experimental control data, supplementing the primary data record.

9.5.4: Success Criteria 2: Data from all cases preserved in the long term

- Data is formatted and packaged in a way that enables both long-term preservation and reuse
- Assessment: Data from all cases is stored and experts evaluate methods and formats as relevant

[92] Accessed 11062017
[93] https://zenodo.org/, https://figshare.com/ and https://dataverse.org/, accessed 11062017
[94] https://www.sa.dk/da/forskning/for-forskere/anmeldelse-aflevering-forskningsdata/, accessed 11062017
[95] https://ckan.org/ and http://www.dspace.org/, accessed 11062017

This success criterion is largely related to the work done in the cases, and we will summarise this in the following table:

Table 9.1: Cases and data stored

Case	**Data/Format**	**Time Range**	**Repository**
LARM	application/JSON	10 years minimum	LOAR/LCAR
Netlab	CDXJ, CSV, JSON, .txt mm	10 years minimum	LOAR/LCAR
Kierkegaard Case	Kierkegaard Normalformat 1	50 years minimum	ERDA (KU) - DATAVERSE (KB)
Local Elections Surveys	DDI Lifecycle	Indefinitely	Danish National Archives
CALPIU	audio and video recordings, CLAN files	Indefinitely	LCAR is a possibility
GALAXY	clinical data from patients in treatment and data from tissue samples and animal experiments	Indefinitely	Danish National Archives/ Zenodo
CAFF terrestrial CBMP	video recordings, geo locator data, eggshell thickness (.xslx), environmental chemistry, etc.	Indefinitely	The Arctic Biodiversity Data Service
DTU Wind energy	Meteorological data: measurement data, computational models, experimental data, etc.	Indefinitely	Zenodo
DTU Space	Geomagnetic data	Indefinitely	Institutional repository at DTU Space. Evaluation ongoing.
Kepler Case	BagIt File Packaging Format	50 years minimum	ERDA (KU) - DATAVERSE (KB) - KASOC (AU)

As can be seen in the table, not all data is packaged and stored. Plans are in place for all the data, however. There are different reasons for data not yet being preserved in a repository. Some cases or projects have been extended and the researchers are still working with the data. Some cases have legal issues that have not yet been resolved. Some cases are waiting for institutional repositories to be established.

In summary, some cases have data preserved in repositories. For the remaining cases, temporary solutions for data have been found and/or plans for long-term preservation are in place for when the projects reach the 'deposit for long-term preservation' step in the lifecycle.

9.5.4.1: LOAR: Library Open Access Repository

The LOAR (Library Open Access Repository) pilot service is based on DSpace and is available from https://loar.statsbiblioteket.dk/xmlui[96]. LOAR is an open data repository established as a service for storing and providing access to Danish research data. It is open to all researchers in Denmark.
The service has the following key aims:

- Make data accessible for review for publications
- Enable researchers to meet requirements for Danish and European grants
- Ensure data privacy and removal of data as appropriate
- Enable reuse of data where appropriate

Researchers who upload data are expected to share the data using Creative Commons licences.

9.5.4.2: LCAR: Library Controlled Access Repository

The proposed LCAR (Library Controlled Access Repository) service is also based on DSpace and is in many respects a twin to LOAR. The key aim of LCAR is to ensure preservation of possibly sensitive research data that requires restricted access.

[96] Accessed 11062017

9.6: Training and marketing toolkits

Developing training material is not an easy task, especially as there is no homogeneity in how institutions address formal data management training. Furthermore, this does not even take the informal training into account that occurs at various levels, e.g. departmental meetings.

Various factors influence the way formal training is conducted:

- Context of the training: Data management can be part of other courses, e.g. part of an overall course on code of conduct or organized as individual courses centred on data management itself.
- Time: Courses in data management, whether or not they are entire courses or part of other courses, can last from around one hour to a half a day or a full day's training.
- Target group: Courses vary in terms of target group from project groups of researchers to Ph.D. students.
- Discipline: Training can be either targeted at a group of people within the same discipline, or it can be targeted at a wider group, e.g. Ph.D. students throughout an entire university.
- Teaching method: Various trainers use various training methods, often determined by time and experience. Some will prioritize time for exercises, while others might stick to traditional "blackboard training" or use e.g. flipped classroom.
- Preparation: Some training will include mandatory preparation, while other training will be available for walk-in without any preparation. Some preparation includes filling out a data management plan before the course, or taking one or more lessons in an e-learning solution like Mantra.

With all this heterogeneity in mind, the working group found it hard to determine what kind of material would be useful for training. It made sense to conclude that one-size-fits-no-one, and that static training material like slideshows centred on various topics would be insufficient and hard to adapt to each institution. Setting up an electronic course environment was not feasible within the allocated time and, with the profile of the participants, would also collide with other national initiatives.

9.6.1: The lifecycle and DMP challenge – a detour

There seems to be a high degree of conformity in the data management community in terms of thinking of data as one single large lump. We can make contracts to deal with data, we make Data Management Plans that describe data, and so on. However, data is not just one clear-cut phenomenon, and neither are the licence terms for data, for example. Data changes over time – that is why it is often collected in datasets. And depending on various factors like time and purpose, the dataset might have a certain role (e.g. designed for publication) and thereby have certain licensing terms that were different from the licensing terms governing the original dataset. Data Management Plans, in particular, do not cope well with this complexity if they try to cover large datasets. The purpose of training should therefore reflect this complexity and enhance the participant's ability to see the

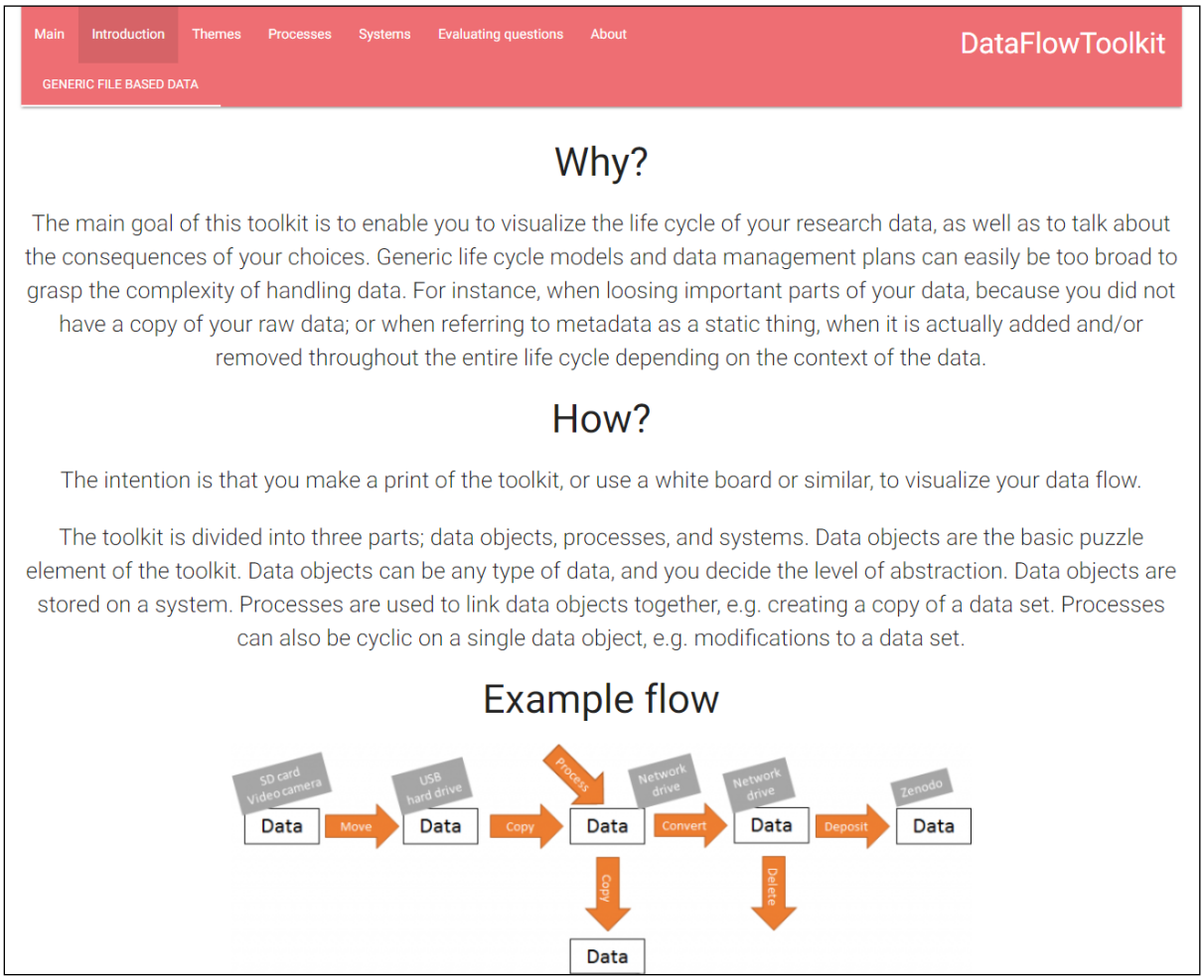
intermediate results of their choices and consequences of their actions on handling data. This is very well aligned with the idea of drawing diagrams to illustrate flows.

9.6.2: Data lifecycle as a flow – back on track

The idea of having a flowchart for data is not new. Flowcharts are used in the industry and elsewhere to illustrate progress and dependencies of processes and to gain insight into changes over time. This facilitates a much more diverse way of thinking about data, and enables insights to be gained into the research data lifecycle. However, the flowchart can be much more complex than a simple lifecycle and may illustrate that different datasets have different lifecycles.

The group then invented the "product" called DataFlowToolkit (abbreviated to DFT), which can be used in teaching and learning about research data management.

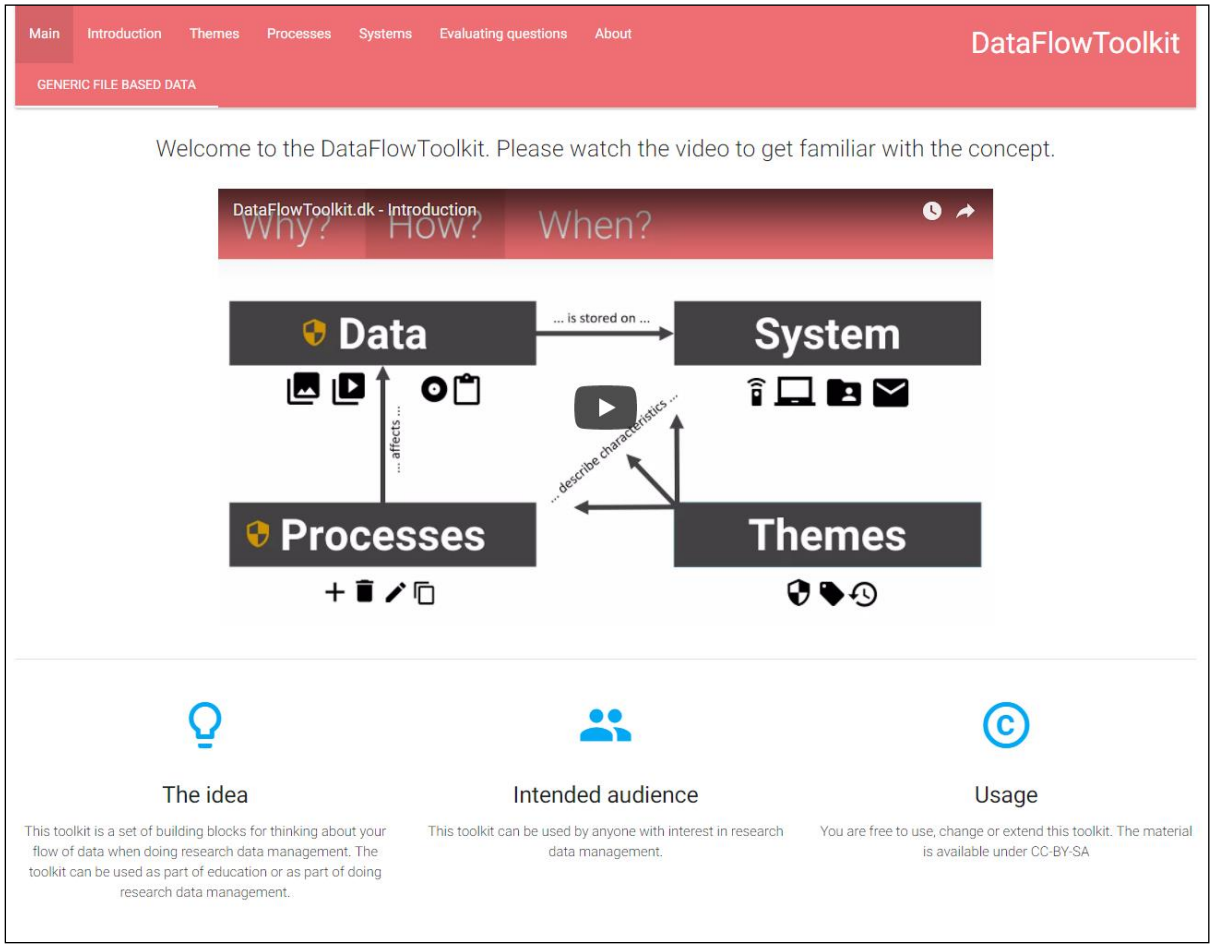
Figure 9.1: Presentation of what Data Flow Toolkit is and why it should be used



DFT is created as a website (www.dataflowtoolkit.dk[97]), but is also a model that can be used to describe the flow of research data. The model consists of three parts:

- Object: An abstract representation that can be used to describe data. The abstraction level is free for the user in terms of whether there should be one or more objects to represent the entire dataset. This is primarily down to whether the properties are the same for all sub-objects.
- Process: Processes are actions that a human or system applies to data. These can include copying a dataset, merging it, etc. A process can be either from one object to another, or a cyclical process that, for example, alters the object itself.
- System/storage: All objects are stored on a system that can range from simple storage systems like USB sticks to more complex database setups or similar.

Figure 9.2: Overview of Data Flow Toolkit

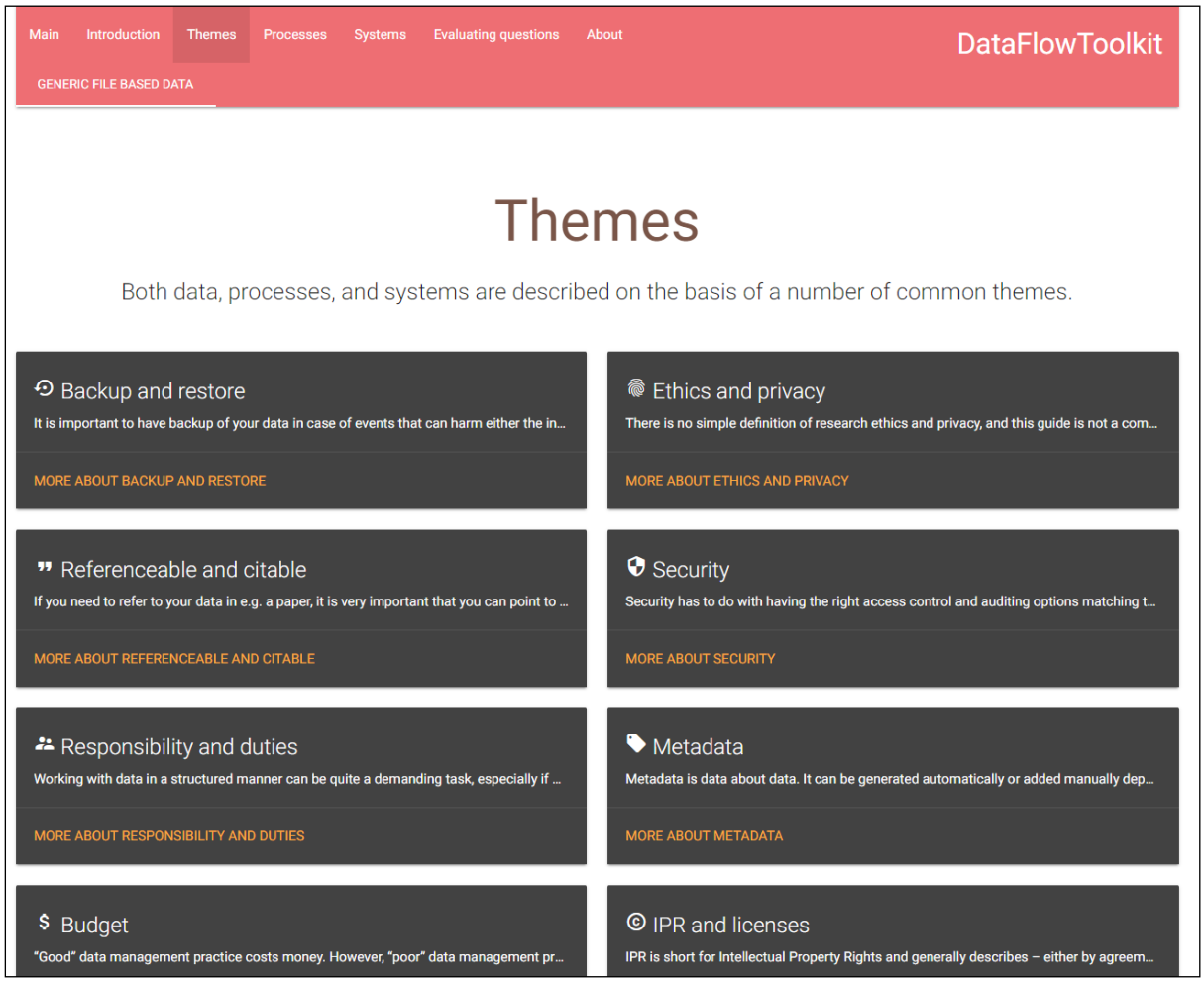


[97] Accessed 11032017

Objects, processes and systems/storage share a number of themes that describe characteristic observation points for each of the three. The current themes are:

- Backup and restore
- Budget
- Ethics and privacy
- File formats
- Governance and management
- Integrity and quality assurance
- IPR and licences
- Metadata
- Referenceable and citable
- Responsibility and duties
- Security

Figure 9.3: The themes as presented online in Data Flow Toolkit:



Not all themes apply to all objects, processes and systems/storage. Please take a look at www.dataflowtoolkit.dk for further details on how these themes are used and described.

When a flow has been created (this will be done by either drawing or using the print functionality for getting the texts), the toolkit also includes a number of questions that can be used to evaluate if the current flow is best suited for, for example, preserving the integrity of data or if data is ready for publication.

9.6.3: DataFlowToolkit for training

DFT can be used in a variety of ways for conducting training. The model – including the type of objects, process and storage/systems – can be used as an underlying basis for understanding the need for doing research data management the right way. It is then highly recommended to use the toolkit for getting the participants to draw their flow and think about each theme throughout their flow – for example to spot the need for different licensing terms for each dataset.

The flow produced can be textualised into a Data Management Plan by each participant if the training is centred around making such a plan.

The toolkit is online now and you can find further information on its technical setup in Appendix 18.

9.7: Sustainability

In order to ensure the services and infrastructures that have been developed during the project beyond the end of the project, a sustainability plan has been developed for some of the services. It has not been possible to make plans for continuing all the services provided in all cases, but most themes have a sustainability plan.

Sustainability has been defined in this project as a continuation of an infrastructure and/or service after the resources of the project ends. A plan for this should describe both the funding opportunities and the organizational responsibility for running the infrastructure and/or service beyond the project period.

The following section will describe the plans for each theme in the amount of available detail. The description will include who will run the service/support/development and who will be responsible for the technical side.

9.7.1: Theme 1: Data Management Planning

The main objective of this theme was to create a Danish installation of the DCC's DMPonline. The software is up and running and includes an H2020 template for Data Management Plans and a beta version of a national Danish template, which has been used in this project in multiple cases. The template designed by DCC was used as a test during a number of workshops for Ph.D. students at UCPH. When the DMiP project ends, the service will continue to offer all researchers in Denmark a platform on which they can create Data Management Plans. The technical operation and server/software maintenance/update will

be handled by DeIC without external funding until the end of 2018. During the same time period (2017-2018), a DEFF-funded project will establish local support and training functions at each university library. The cost of the project is DKK 3.28 million, of which DKK 2.248 million is funded by DEFF.

9.7.2: Theme 2: Data capture

The goal of this theme was to try to create an infrastructure, e.g. a repository, for the data from each case. At the University Library of Copenhagen, a Dataverse body has been set up that holds data from the KASOC, Kepler and Kierkegaard cases. A Dataverse service is also available for the Danish National Archives on the Royal Library servers. The service still remains to be tested, but this is planned in the near future.

The current installation of Dataverse is a proof-of-concept and shows how it could be run by multiple partners in the future. The storage part is now handled by the Niels Bohr Institute (which also runs the ERDA Electronic Research Data Archive for the science department at UCPH), whilst the repository (including user interface, APIs, metadata database and search engine) could be handled by a different partner, e.g. the library. The final design of the service is yet to be settled, but it is clear from the work with the actual data from the cases in the project that the needs of the various institutions and research areas are very different, which poses a challenge to the service model. However, Dataverse has proven versatile and able to function both in local/specialized and/or in national/generalized contexts, so it should be possible to find a suitable model. This and previous projects have shown that there is a strong potential for the university, the university library and/or other partners to cooperate on a service like this. It will not be possible to find a model within the scope of this project, but the work towards agreement will continue.

9.7.3: Theme 3: Data identification

The objective of this theme was to make sure that the publishable data from the cases was put into repositories that were able to connect to the DataCite service at DTU, so that each dataset was described using the DataCite Metadata Schema. The DataCite service is not a product of this project, so sustainability of this particular service is not described as such, but there are no indications of a discontinuation in the foreseeable future by the minting service at DTU. It was also a recommendation of this theme that The Danish National Research Database starts to use the DataCite MDS when they begin including datasets in their database.

9.7.4: Theme 4: Long-term preservation

At The Royal Danish Library in Aarhus a long term preservation service was established in the shape of the LOAR Open Access Repository based on DSpace. A plan for the continuation of the repository has been outlined which states that the organizational responsibility lies with The Royal Library, Aarhus, which also covers the associated costs.

Based on experience with the cases in the project, the need for a service that handles data that requires controlled access has been identified. As a result, the Royal Danish Library in Aarhus is currently investigating to what extent DSpace serves this purpose.

9.7.5: Theme 5: Marketing and training

This theme was focused on creating material for training researchers in proper data management with a starting point in their own project – on a 1-on-1 basis. The material has been developed and is feeding into another project sponsored by DEFF: Data Management Guide. The toolkit is technically a simple setup and is usable in both an online and paper version. The toolkit is available on this website: https://dataflowtoolkit.dk/[98]

The toolkit is presently hosted on a university library server, but external partners are being sought for further development and hosting of the tool. Since the cost of running the service is relatively low, this is a viable temporary solution until one or more partners are found. If an external partner is to edit the content of the toolkit, a CMS system has to be used. As for the content of the toolkit, the project recommends that an editorial group is established to run development.

9.7.6: Cases

In all cases, the project members of Data Management in Practice have helped researchers meet the different data-related challenges of their research project: Data Management Planning, advice on long-term preservation (formats, repositories), storage, metadata, publication/sharing of data and so on. The types of assistance required have been different from case to case, but Data Management Planning has been a common denominator in the cases, indicating that DMP assistance will be a prevailing need in the future.

The researchers in the cases have learned a lot about data management in general – as have the project members. The interaction with the researchers will certainly have led to a greater sense of confidence in many aspects of managing research data on both sides. This new knowledge and awareness of tools, infrastructures and services will live on after the project in all participants. When it comes to scaling services up from one case to cover the whole university, the number of man-hours required in the cases will without doubt mean that a lot more human resources will be necessary to deliver the same kind and level of service to the entire institution. The focus on data management from funders, politicians, institutions and publishers will make the need for infrastructures – national, local, or both – services, training and advice from all parts of the institutions (including IT departments, legal

[98] Accessed 11032017

departments, libraries, research support units and from archives) more and more pressing. It is clear that the area of data management is ever more demanding and requires a collective effort on a national, institutional and personal scale in order to deliver data and other documentation of research to the world in a findable, accessible, interoperable and re-usable way.

One example of this is the creation of fora for data management, both at national and institutional level. At DTU, SDU and UCPH, groups of researchers organized around the topic of data management have also been established. In recognition of the fact that data management is a collective project, a group called the SDU DM Forum was formed early last year. This group includes members from the IT and legal departments, the library and from the research support units supporting researchers in the pre-grant phase. The creation of this group has meant that through a number of cases we have demonstrated that there is a need for help and that we have been able to help. It is becoming clear that if we are to meet the demands, we have to have more resources, both in terms of funding for people with the right competencies, but also for the purchase or development of the necessary infrastructures and services. As rules and policies begin to demand Data Management Plans, there will be an increasing need for information, teaching and also assistance in reviewing and archiving of Data Management Plans. It is our hope in the SDU DM Forum that through our cases we can convince university management that this area needs greater priority.

9.8: Conclusion of the thematic section

The six themes above show that the cases described in sections 3 to 7 have given valuable insight into the complexity of the practical side of research data management. In general, the different activities regarding the themes met the requirements that were to be addressed by the project.

As a result of the first theme – Data Management Planning – a test version of DMPonline was established, which, in turn, led to an operating version of the system. As part of the project, two separate versions of a national template were defined.

Regarding the second theme – Data capture, storage and documentation – there are no coordinated or common results. This is due to the fact that the different projects used different infrastructures and thus it would be rather difficult to define a common infrastructure to handle the needs of different research projects from all scientific areas.

The third theme – Data identification, citation and discovery – did not cover all criteria, since the project required that the different cases should publish data placed in institutional repositories, which was not the case. Based on the experience gained from the cases, however, it was still possible to define infrastructures and services necessary to make research data findable and citable, including DOIs from DataCite.

The fourth theme – Select and deposit for long-term preservation – met the different criteria, and the cases gave an insight into how to select and store data. The cases also made it possible to compare different repositories, as a result of which the Royal Library, Aarhus, established an open access data repository: Library Open Access Repository (LOAR). The

project group also assessed PURE as a possible institutional repository. It was deemed that PURE has many – but not all – of the features needed to be used as an institutional research repository.

The fifth theme – Training and marketing toolkits – developed a training tool which makes it possible to help researchers to define how to handle the individual stages of the data lifecycle. This tool was based on experience from the activities in the different cases.

The sixth and final theme of this project was Sustainability. This theme should address how and if services could and should continue after the termination of project. As can be read in section 9.7, some of the services will be maintained, whilst some are currently still being assessed.

10: Evaluation of the project by the international experts

In order to ensure an evaluation based on international experience, it was decided to assign two internationally recognized experts to evaluate the project in its entirety. According to the project application, from the start of the project the experts should be given access to the internal workflow of the project via the project site[99] and the material shared by the project participants, as well as attending the closing conference[100] and, based on this, evaluating the project.

This conference was held in Copenhagen at the Royal Danish Library on 30th March 2017. The conference programme reflected the division between cases and themes. The cases were presented first, followed by the themes in the afternoon. In order to give the evaluators updated information on the different sub-projects, they received an early version of this report a couple of days before the conference. Some of the comments made by the experts are available online[101].

The two experts found that the project addressed many of the challenges related to the area of research data management. Furthermore, the project convincingly documented the problems through the case-based approach. This result was produced by the variety of processes and types of data in the project. The data originated from all subject areas and the activities spanned every phase of the data lifecycle. In addition, the cases involved questions on sensitive data, sharing and findability of data.

The experts identified a number of areas which the project had not covered. Among these were how to handle the software used to create the data and software developed in a project. Another issue, which according to the evaluators could have been addressed more comprehensively, is the licensing of software. For example, does the researcher use Creative Commons[102] or other solutions? The experts would also prefer a closer investigation of the consequences for publishers if datasets were shared on public servers. Finally, the project did not address the question of data citation career incentives for researchers. However, despite the issues that had not been addressed, the experts were impressed by the results of the work of the project group.

The experts suggested that the work of the DMiP project should result in the formulation of a Danish "Vision for Data Management". This vision describes the "why", "what" and "how" of the different topics covered by the project and should specify the stakeholders in the area and their individual roles[103]. The experts suggested the use of the Dutch National Plan

[99] https://sbprojects.statsbiblioteket.dk/display/DAT, accessed 11072017
[100] Andersen 2014, page 6.
[101] https://openworking.wordpress.com/2017/03/30/feedback-for-danish-research-data-management-project-dimp/, accessed 11072017
[102] https://creativecommons.org/, accessed 11072017
[103] https://openworking.wordpress.com/2017/03/30/feedback-for-danish-research-data-management-project-dimp/, slide 7-13, accessed 11072017

for Open Science[104]. After the conference, the project group decided that it would not try to formulate a national vision for open science. This was not due to lack of interest, but because it would be outside the mandate of the project group.

10.1: About the two experts

The two experts were Maria Johnsson and Alastair Dunning.

Maria Johnsson is a librarian at the University Library at Lund University. Maria works at the Department of Scholarly Communication, where she has been working with data management since 2014. She has published several articles on data management and is co-author of a report titled "Research Libraries and Research Data Management within the Humanities and Social Sciences[105]".

Alastair Dunning is head of Research Data Support at TU Delft and Head of the 4TU.Centre for Research Data, the Dutch data archive for the technical sciences. Alastair has been working with digitization and data management since 1999 and has previously worked at the Arts and Humanities Data Service, JISC and Europeana. He has published articles on digitization and data management and his blog his available at Available Online[106].

[104] https://repository.tudelft.nl/islandora/object/uuid:9e9fa82e-06c1-4d0d-9e20-5620259a6c65?collection=research, accessed 11072017
[105] Åhlfeldt and Johnsson, 2015
[106] https://availableonline.wordpress.com/, accessed 11072017

11: Conclusions

The original project application (see Appendix 1) listed the goals and criteria which the project had to achieve and meet during the project period. As shown in the different descriptions of cases and themes, most of these have been met. In this section, we will review the results.

The first criterion was that a Danish infrastructure should be established covering the entire research data lifecycle. This should be based on international best practice and should be assessed by international experts. This has been partially met, with the DMPonline server and LOAR servers having been established as a result of the project. As mentioned in section 9.3, the variety of data in the cases made it impossible to define a common structure. The chosen technical solutions still meet international standards. As an example, one of the templates used in DMPonline is the H2020 template. As can been seen in section 10, the two experts also approved the results of the project group.

The second requirement was that the project should demonstrate that the research libraries have a role to play in research data management. The project should actively ensure that the competences of the libraries have developed and can be regarded as part of the national Danish solutions. It should also be evaluated how the results of the project are integrated into the work of the National Forum for Research Data Management[107].

The project shows that research libraries and archives can play different roles according to the institutional framework and functions of the individual library and archive. Some institutions operate services such as the Netarchive, The Danish Data Archive and the Cultural Heritage Cluster. These institutions and their employees will have other job functions than librarians working at university libraries, where the access to research data is not necessarily based within the university library. However, these librarians can offer advice as to where and how to find data, hold courses on research data management and much more.

The project itself has developed the knowhow of the involved staff members and their professional network. This knowledge stems from operating the different services. For example, the project group developed a proposal for a national research data management planning template to be used in the DMPonline installation (see sections 9.1 and 9.2).

The National Forum for Research Data Management and the members of the project group have been in close contact during the project. Some members of the group were members of the forum, whilst from September 2017 Anders Sparre Conrad became Chairman of the National Forum for Research Data Management. As mentioned in section 3.4, Aarhus University, the forum and a member of the project group formulated a model agreement on Data Management in Cooperative Research Projects.

With regard to Data Management Planning, the project had to establish a Danish version of DMPonline, which should be available to Danish researchers and adapted to Danish

[107] https://www.deic.dk/da/datamanagement/DM_forum (in Danish only), accessed 11072017

conditions. This was achieved by the installation of a test server for the project, followed by a separate DMPonline server. As mentioned in section 9.1, a Danish national template for data management planning was developed. In general, the involved researchers considered DMPonline to be a useful tool.

With regard to data capture, storage and documentation, the criteria that the project had to meet were that new or existing data repositories should be available and that technology for capturing and storing data should be implemented, thus providing a direct link from the research process to a data repository. This criterion has only been partially met. However, the variety of cases and data was so great that further progress in this area was not possible. See section 9.3.

As regards data identification, citation and discovery, the project had to meet several criteria. Data from the different cases had to be stored and provided with persistent IDs. It was a condition that researchers would receive relevant academic credit for the research producing the specific dataset(s). The use of ORCID IDs should help to ensure this. The data should also be findable in a national database. As is described in section 9.4, the data from the different cases could not deliver data for the theme. Accordingly, the success criteria had to be changed and, instead, the project had to focus on a series of recommendations:

- Dataset must provide metadata based on DataCites Metadata Schema version 4.x
- Datasets should be provided with a DOI
- Researchers should use ORCIDS

As regards select and deposit for long-term preservation, the project was designed to create relevant documentation based on best practice and describe criteria for how data that is to be preserved is selected. The data from all cases should be formatted so that the data could be reused. Based on the cases, we recommend that researchers deposit data in discipline-specific or specialist repositories. In these repositories, we expect that data is more likely to be preserved correctly and be discovered by fellow researchers and specialists in the field. However, repositories without a specific disciplinary or institutional connection, like LOAR, can of course also be used. As can be seen in section 9.5.4., not all the cases have deposited data in repositories for long-term preservation. The different cases were in different phases of the data lifecycle, which has created an outcome where relatively few datasets were stored as a direct result of the project. However, as a result of the project, a repository, LOAR, was created for long-term preservation.

With regard to training and marketing toolkits, the project should result in materials which can be used for training. In addition, plans for outreach in selected research groups should be created. As mentioned above, the Data Flow Toolkit was established. In all the cases an intensive dialogue between researchers and project participants has taken place. This has ensured a high level of training of the involved researchers and project group members. The experience gained is currently being used in a project to ensure that researchers from Danish universities use the national DMPonline. The project has fulfilled the criteria within this theme.

In relation to the last of the themes, sustainability, the aim was to develop business models for relevant services. Some of the services mentioned above, e.g. DMPonline, have received additional funding. The cases, and the experience drawn from these, have accumulated knowledge which can be used in relation to local and national networks and will make it easier to ensure that new services can be established and existing services are improved. Nevertheless, in order to ensure permanent funding of services, to some extent payment for use will always be necessary.

In general, the criteria for success have been met whenever possible with regard to practical, technical and financial conditions. In addition, the project has given a clear understanding of the fact that research data is so diverse in nature that the concept of one-size-fits-all makes no sense when planning for the provision of the necessary services.

12: References

Andersen, Bjarne. 2014. *Data management i praksis*, Statsbiblioteket. See Appendix 1

Andersen, Jesper Steen, Bente Larsen og Jacob Tøgersen (Ed.). 2013. *LARM Audio Research Archive*: København, Københavns Universitet, https://larm.sites.ku.dk/files/2014/01/LARM-BOG_tryk_indhold.pdfPDF, accessed 10262017

Conrad, Anders, Micheal Svendsen and Rasmus Handberg. 2016. Back-up to the Future: Creating the Kepler/K2 long-term "living archive".
Poster at the SPACEINN & HELAS8 Conference in 2016. Url. : http://www.iastro.pt/research/conferences/spacetk16/posterDetails.html?abs=147, accessed 11062017.

Conrad, Anders, Micheal Svendsen. 2017. *Reuse for Research. Curating Astrophysical Datasets for Future Researchers*,
Practice Paper at the IDCC17. https://www.slideshare.net/MichaelSvendsen1/reuse-for-research-presentation-idcc17, accessed 11062017.

Conrad, Anders, Micheal Svendsen and Rasmus Handberg. 2017. *Reuse for Research. Curating Astrophysical Datasets for Future Researchers*.
Presentation at Parallel B4: Curation & Reuse, IDCC17, 22/2 2017. Url to presentation: https://docs.google.com/presentation/d/1Gcki1VhxdYAA-GaBpUzt-ofDxGIfXF1VbNrMG7LXK6Q/htmlpresent, accessed 11062017 and Url. of conference http://www.dcc.ac.uk/events/idcc17/programme, accessed 11062017.

DataCite Metadata Working Group. 2016. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data*. Version 4.0. DataCite e.V. http://doi.org/10.5438/0012.

Handberg, Rasmus, Günter Houdek, Jørgen Christensen-Dalsgaard, Anders Sparre Conrad and Michael Svendsen. 2017. *Long term KASOC Archive*. http://www.spaceinn.eu/wp-content/uploads/2017/01/D3.15_SpaceINN_Deliverable_KASCO_Archive.pdf, accessed 11062017.

Jensen, Erik Granly, Jacob Kreutzfeldt, Morten Michelsen og Erik Svendsen (Ed.). 2015. *Radioverdner*: Aarhus, Aarhus Universitetsforlag, 2015.

Jurik, Bolette and Eld Zierau.2017. *Data management of web archive research data*. Paper presented at the “Researchers, practitioners and their use of the archived web” conference, 2017. DOI: 10.14296/resaw.0002.
https://archivedweb.blogs.sas.ac.uk/files/2017/06/RESAW2017-JurikZierau-Data_management_of_web_archive_research_data.pdf, accessed 11032017

Kruse, Filip and Jesper Boserup Thestrup. 2014: *Research Libraries' New Role in Research Data Management, Current Trends and Visions in Denmark*. LIBER Quarterly, 23.4:310 – 335 https://www.liberquarterly.eu/articles/10.18352/lq.9173/; accessed 11012017

Kunze, J., J. Littman, L. Madden, E. Summers, A. Boyko and B. Vargas. 2016. *The BagIt File Packaging Format (V0.97)*. https://tools.ietf.org/html/draft-kunze-bagit-14. Accessed 11032017.

Larsen, Asger Væring, Sacha Zurcher, Mikkel Hvidtfeldt Andersen, Jesper Boserup Thestrup, Søren Ærendahl Mikkelsen, Mogens Sandfær and Anders Conrad. 2014. *Fælles Infrastruktur for Forskningsdata*. Copenhagen: DEFF.

Thestrup, Jesper B., Filip Kruse, Lars Nondal, Bertil F. Dorch, Mikkel Andersen, Niels Jørgen Blaabjerg, Thea Drachen, Jeannette Ekstrøm, Mikael Elbæk, Ole Ellegaard, Karsten Kryger, Asger V. Larsen, Diba Markus, Anna Mette Morthorst, Erik Sonne, Sacha Zurcher and Ellen V. Knudsen. 2012. *Forvaltning af forskningsdata i Danmark. En undersøgelse af danske universiteters og forskningsråds praksisser, politik og visioner for lagring, langtidsbevaring, tilgængeliggørelse og deling af forskningsdata*. DEFF.

van Wezenbeek, W.J.S.M, H.J.J. Touwen, A.M.C. Versteeg and Astrid van Wesenbeeck. 2017. Nationaal plan open science. Delft. https://repository.tudelft.nl/islandora/object/uuid:9e9fa82e-06c1-4d0d-9e20-5620259a6c65?collection=research, accessed 11062017

Wilkinson, Mark D. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18, http://www.nature.com/articles/sdata201618.pdfPDF, accessed 11032017.

Zierau, Eld, Caroline Nyvang, and Thomas Hvid Kromann. 2016. *Persistent Web References – Best Practices and New Suggestions* in Proceedings of the 13th International Conference on Preservation of Digital Objects (iPres), pp. 237–46. http://www.ipres2016.ch/frontend/organizers/media/iPRES2016/_PDF/IPR16.Proceedings_4_Web_Broschuere_Link.pdf, accessed 11032017

Åhlfeldt, Joahn and Maria Johnsson. 2015. *Research Libraries and Research Data Management within the Humanities and Social Sciences*, Lund University. http://portal.research.lu.se/ws/files/6286782/5050466.pdf, accessed 11062017.