LIBER QUARTERLY
THE JOURNAL OF THE ASSOCIATION OF EUROPEAN RESEARCH LIBRARIES

Vol. 23, no. 4 (2014) 310–335 | ISSN: 1435-5205 | e-ISSN: 2213-056X

Research libraries' new role in research data management, current trends and visions in Denmark

Filip Kruse

The State and University Library, University Library Division,
Aarhus, Denmark
fkr@statsbiblioteket.dk

Jesper Boserup Thestrup

The State and University Library, University Library Division,
Aarhus, Denmark
jbt@statsbiblioteket.dk

Abstract

The amount of research data is growing constantly, due to new technology with new potentials for collecting and analysing both digital data and research objects. This growth creates a demand for a coherent IT-infrastructure. Such an infrastructure must be able to provide facilities for storage, preservation and a more open access to data in order to fulfil the demands from the researchers themselves, the research councils and research foundations.

This paper presents the findings of a research project carried out under the auspices of DEFF (Danmarks Elektroniske Fag- og Forskningsbibliotek — Denmark's Electronic Research Library)[1] to analyse how the Danish universities store, preserve and provide access to research data. It shows that they do not have a common IT-infrastructure for research data manage-

Igitur publishing | http://liber.library.uu.nl/ | URN:NBN:NL:UI:10-1-116068

ment. This paper describes the various paths chosen by individual universities and research institutions, and the background for their strategies of research data management. Among the main reasons for the uneven practices are the lack of a national policy in this field, the different scientific traditions and cultures and the differences in the use and organization of IT-services.

This development contains several perspectives that are of particular relevance to research libraries. As they already curate digital collections and are active in establishing web archives, the research libraries become involved in research and dissemination of knowledge in new ways. This paper gives examples of how The State and University Library's services facilitate research data management with special regard to digitization of research objects, storage, preservation and sharing of research data.

This paper concludes that the experience and skills of research libraries make the libraries important partners in a research data management infrastructure.

Key Words: research data management; university library; national cultural heritage

1. Introduction

The focus of this paper is on how national libraries and university or research libraries have a new and important part to play in research data management. First, in Section 2, we outline the concept and changing reality of research data and we make a brief presentation of the results of a DEFF research project (Danmarks Elektroniske Fag- og Forskningsbibliotek — Denmark's Electronic Research Library) that charted the field of how research data management is carried out in the universities of Denmark. The activities connected with research data management such as providing facilities for data reuse and data sharing change the concept and reality of research data. We discuss this in Section 3. As part of its obligation to preserve the national cultural heritage, The State and University Library is already providing an e-infrastructure for researchers' access to its radio, television and web archives. In Section 4 we show how the cooperation between researchers and library staff is a precondition for developing and providing the right facilities for access to, and use of, data.

2. The DEFF project Research Data Management in Denmark

2.1 Aim of the project

In 2013, a project group finished a project initiated and financed by DEFF with the publishing of the report *Forvaltning af forskningsdata I Danmark* (Research Data Management in Denmark) (Thestrup et al., 2013). The aim of this project was to gain knowledge on a subject hitherto unexplored in Denmark, namely how the Danish universities' work on research data management is organized and how it is carried out in practice, especially with regard to data storage, data preservation and data sharing or other forms of making data more openly accessible.[2] All Danish universities participated in the survey.

Closely connected to the aim of establishing an overview in the form of a catalogue of knowledge of the universities' current activities and future plans in this field, it was also important to supplement this with the most significant international experiences in this field. In order to complete the picture of the current situation in the field of research data management in Denmark, it was also necessary to explore the leading Danish research councils' and foundations' viewpoints towards increased data accessibility. The overall vision for the project was to increase the attention and the interest of decision makers in this field, on the political level as well as on the university level.

2.2 Methods employed

The study is explorative in its design and its focus is on the present state of research data management at the level of policy and strategy, and not on the researchers' specific activities.

The project group carrying out the survey consisted of participants from all the universities' libraries: The Royal Library,[3] Copenhagen Business School Library,[4] Aarhus University Library,[5] State and University Library,[6] DTU Library,[7] University of Southern Denmark Library,[8] Aalborg University Library[9] and Roskilde University Library.[10]

The survey was carried out as a series of personal interviews with top level managers from the universities, university libraries and university IT-departments. The interview topics[11] covered the individual university's plans and activities with regard to data storage, data preservation, data archives, enhancing the researchers' awareness of data preservation, visibility and accessibility, and whether the university's exposure of research data is considered to be a part of its branding strategy. The personal interview was chosen, as the data-collecting technique since it — in contrast to the ordinary questionnaire-based survey — permits dialogic interaction such as follow-up questions to the interviewee's answers. Regarding the research councils and foundations, a slightly altered questionnaire was issued based on the interview guide, and it was distributed by e-mail.

Our intention by the first three seemingly overlapping topics was to ensure that the interview would cover as wide a range as possible of the various activities undertaken by universities to secure research data and to make them accessible during and after the research process.[12] Likewise, we used the term 'research data' in the interview guide in a broad sense, as the aim was that the interviewees should give as much information as possible on their activities in this field. Data storage is essentially data backup. We know from this and other surveys in the field (Sørensen et al., 2009; Pattenden-Fail et al., 2010), that researchers' backup actions span from the use of memory sticks, over external hard drives and mail systems to the use of various cloud services. General characteristics of data storage or data backup are that the data are not meant to be accessed by others than the creator(s) and consequently they are not issued with metadata or other means of making them searchable. Further, changes will be made to the data during the research process and the first data collected will not necessarily be saved.

2.3 Concepts and definitions

Data preservation, in contrast to data backup, means securing permanent access to the original research data, and, as a rule, data from the finished research project. General characteristics of data-preservation actions are that the data are accessible to others for verification — such as datasets as part of scientific publications — or for sharing or collaboration within the scientific community. Consequently, the data must be organized and made searchable within institutional settings such as universities' repositories or data archives, or other data archives. Conditions for access to data and for reuse are formally regulated. Long-term preservation of the original data and protection of sensitive data are vital characteristics of data-preservation actions.

Data archives[13] are institutions for data preservation; they are mostly subject-specific, organized on an institutional, national or international level. National data archives are frequently older institutions than, e.g., institutional repositories.

3. Main results of the project

3.1 Data storage

Apart from a few exceptions, Danish universities, as a rule, provide researchers only with general access to the university computer drives, and not with special facilities for data storage. The researchers themselves decide upon which type of storage they want to use, provided that it is secure. The responsibility often lies with the university department. The reasons for this policy, as explained by the respondents, are that data storage, preservation and sharing are not considered to be as important as the publication of research results. Another reason is that considerable uncertainty exists regarding the definition of research data and how to store or preserve them securely. This is reflected in the absence of centralized responsibility in favour of a transfer to departments and individual researchers, in keeping with a sceptical attitude towards a uniform (national) policy and common technological solutions — a so-called 'one-size-fits-all' arrangement. The multitude of different data types and specific needs in relation to storage are also arguments in favour of this.

The research councils and foundations make no specific demands on data storage for the research projects they have funded but expect legal and academic standards to be followed.

3.2 Data preservation

This activity is, in general, considered too resource-demanding for the universities to undertake, and with a 10+ years' time scale, it is not regarded as a natural task for the universities. Two universities have storage facilities which can also be used for long-term data preservation. There are only a few examples of research data being preserved by universities. Again, the different data types and preservation needs, together with the lack of legal obligations, are arguments for this somewhat hesitant policy. Correspondingly, not much attention from the managerial level is given to the potentials of making research data accessible or shareable. The research councils and funds have no formulated policy for data preservation but regard it, to some extent, as the responsibility of the institutions.

3.3 Data archives

The universities favour a national solution consisting of several data archives as a supplement to, and an expansion of, the existing archives. One university favours international data archives based on its specific experiences in 'big science'. The arguments for the national solution are the qualitative and quantitative differences in needs, such as data types, academic traditions and cultures and size of datasets, and also research cooperation between universities, common principles for access to data and financial costs. A national solution also calls for cooperation between the existing data archives. A few universities are, at present, considering establishing their own archives but are awaiting results from more detailed analyses.

The research councils and foundation funds are, in principle, in favour of making data accessible or shareable as this will improve utilization of the resources. This issue must be deliberated in the context of a national research infrastructure, the existing data archives, a coherent definition of research data, standardization of metadata and legal matters in connection with data sharing.

3.4 Researchers' awareness and research data as university branding

The majority of universities is already carrying out campaigns or other activities — or is planning to do so — with the aim of increasing the researchers' awareness of the potentials of preservation and accessibility of research data. All the universities in this study emphasize the importance of publication of research results as part of their strategy, and apart from new ground-breaking data, publications are considered more important than research data, as elements of strategic branding.

3.5 Perspectives

A general result of the survey is that activities in the field of research data storage, preservation, archiving and sharing vary from university to university. This is probably due to differences in institutional and organizational structure, in size and in academic profile. Further, it appears that the universities do not seem to be very active in the field. How should we interpret this?

One way of seeing this is that the universities take a position best described as 'wait-and-see' on the issue of research data. But at the same time, they express concern in relation to the researchers' needs and do not want to exert pressure on them, e.g., in the form of mandatory rules for data preservation. Decentralized solutions are preferred, as they are seen as potentially more suitable to local needs and available resources. This consideration could also be seen reflected in the emphasis expressed on issues related to different needs, different data types, various sizes of datasets, different academic cultures, etc. The need for decentralized solutions can be seen in the cases we describe in Section 4. Researchers were involved in designing the software for handling the research objects, in order to meet their specific needs.

In addition, the launching of various local projects could be regarded as attempts to explore this mainly uncharted territory. A supplementary explanation could be that the universities are unsure of how they should handle the task, e.g., because of lack of experience in the matter. Funding is a separate concern, and publication of research results is a prerequisite for financial support for research projects, while data preservation and sharing are not. At the same time, both universities and research councils and foundations emphasize — in our opinion rightly — the importance of a clear definition of research data, standardization of metadata, clarification of central legal matters in relation to data sharing and a clear division of labour in relation to the data archives.

A quote from the questionnaire of one of the Danish research funds:

> "The fund believes that data archives should be national or international, not institutional. The recent fusions, divisions and closing down of e.g., university departments demonstrates that the individual university's infrastructure has a relative short life and is ill-suited for long-term preservation of research data."

In our opinion, the universities' position is clearly understandable, as they attempt to navigate in a space characterized by different needs in the various fields of research, absence of clear definition of the problem's nature, lack of financial incentives, lack of a national policy on the issue and perhaps also lack of experience in data preservation, use of metadata and sharing of research objects. The national and research libraries already have a stable organizational position and are currently involved in various projects here. A few examples to illustrate this point:[14] recording and preservation of research data in two universities (ORBIT, Danish Technical University, RUDAR, Roskilde University), software development (DATAVERSE, The Royal Library), archiving the Danish internet (Netarchive, The State and University Library) and archiving Danish radio broadcasts (LARM, The State and University Library). The libraries have skills and experience in generating metadata for research objects and datasets, in preservation and in facilitating access to and use of data. Therefore, the libraries are the researchers' natural collaborators in developing and implementing the systems necessary for the management of research data.

Based on the findings of the survey, the project group has submitted several recommendations, among them, that a national policy, as well as a policy for the universities for preservation and sharing of research data, is formulated and that the necessary e-infrastructures are established.

A study of European and international practices in research data management[15] shows some interesting different findings. Here, funders and publishers are found to be the main drivers for work on data management plans and thus also contribute to a national policy in the field. In Denmark, the research councils and foundations are not particularly active, but seem to be waiting for a national policy, rather than developing policies of their own. The Danish universities are in the process of implementing research data management activities, but both the universities and the research funders seem yet to be lagging behind compared to the European and international development, but e.g., DeIC's activities indicate a growing awareness of the issue.

3.6 Same play, new actors — an update

Since the DEFF project was completed, several different institutions in Denmark have been active in developing infrastructures for research data management.

Various activities have been carried out at the individual university level in order to ensure internal processes regarding management of research data. At the national level, DeIC (the Danish E-Infrastructure Cooperation)[16] and DEFF are working to evaluate different services involved in data management from application for grants to preservation and sharing of data. DeIC was founded in 2012, with the primary task to "support Denmark as an e-Science nation through delivery of e-infrastructures." DeIC was established as a merger between Forskningsnettet (the Danish Research Network) and the Danish Center for Scientific Computing (DCSC).

DeIC has formulated six strategic goals. One of them is to "Coordinate solutions concerning Data Management and large datasets"[17]:

> "The amount of research data rises continuously as do the demands for long term preservation. The possibility to reuse and share data across research groups is also a criterion. The institutions' interest in and need for solutions are urgent. Stakeholders: the universities, the research libraries and other research institutions." (DeIC, 2012)

One of the projects in which DeIC is involved is FIF (Fælles Infrastruktur for Forskningsdata — Common Infrastructure for Research Data). The aim of this project is to develop a Danish infrastructure for data management (Christensen-Dalsgaard, 2013). The project, which is based on a grant from DEFF, involves resources from DeIC as well. Since DEFF represents libraries and DeIC represents the universities, the project demonstrates that research libraries are regarded as natural partners by the universities, regarding management of research data.

Parallel to this, DeIC has initiated a process to ensure that a national strategy for data management is formulated. In order to do that, DeIC has contacted the Danish universities, the Danish research councils, The State and University Library, The Royal Library, DDA (the Danish Data Archive), DEFF and a research infrastructure project called DigHumLab (Digital Humanities Laboratory[18]). DeIC wants to formulate a policy which can be approved by the relevant institutions and by the Ministry of Science, Innovation and Higher Education[19] (DeIC, 2013).

Later we shall show that some of the services provided for researchers by The State and University Library in connection with its obligation to the preservation of the national cultural heritage, already facilitate the reuse and sharing of research data. Facilities for sharing research data or other forms of improving data accessibility are an important part of the requirements of data management systems (see, e.g., Higgins, 2012). Further, we shall discuss future perspectives generated by these services for the role of research libraries. But before we turn our attention to these matters, we shall consider how the concept and reality of research data has changed, and discuss the possible implications of this development.[20]

4. From research data to research objects and vice versa

4.1 Research data

In order to provide a conceptual framework for further analysis it is necessary briefly to consider the concepts 'research results', 'research objects' and 'research data'. Several initial definitions are provided by the Data Information Specialist Committee — UK:

> "…research data, that which is collected, observed, or created, for purposes of analysing to produce original research results. This differs from what is commonly called research outputs, which are the peer reviewed, published papers/articles/books/presentations that are produced as a result of data analysis. Research data may be created in tabular, statistical, numeric, geospatial, image, multimedia or other formats." (http://www.disc-uk.org/qanda.html)

Research results are outputs of the research process, which has its focus on a specific research object. Research data are thus always data 'on' a specific research object, for instance, Stone Age tools, solar flares or political extremism in Europe. Research data can be seen as records of the research activity, i.e., that which is created by the activity and preserved for use or reference in the future. Consequently, the records must be authentic, reliable, usable, complete and unaltered (Higgins, 2012, p. 20). Research data can be primary or secondary.

> "Primary data are data that are collected for the specific research problem at hand, using procedures that fit the research problem best. On every occasion that primary data are collected, new data are added to the existing store of social knowledge. Increasingly, this material created by other

> researchers is made available for reuse by the general research community; it is then called secondary data." (Hox & Boeije, 2005, p. 593)

This distinction is based on difference in the purpose of the data collection and consequently also on its relation to the research process: Whereas primary data are collected with the purpose to contribute to solving a specific research problem, secondary data are collected with different research purposes than the one for which they are initially used.[21] The researcher using secondary data "by an act of *abstraction* uses questions originally employed to indicate one entity to illuminate other aspects that a former analyst did not have in mind at all." (Hyman, 1972, p. 37)

Primary data can be collected through methods such as experiments, clinical tests, qualitative interviews and surveys, depending on the specific research project and the area of science to which it belongs. Social science researchers often either produce their own data through methods such as surveys, interviews and field studies, or they use data from statistical records. Science data are also produced by the researchers, often by observations, experiments and computer models, or they are drawn from data archives, e.g., researchers in astronomy and genomics have for some time shared their data in common archives (Borgman, 2012). Humanities data "most often are drawn from records of human culture, whether archival materials, published documents, or artefacts" (op. cit. p. 1061).

Primary and secondary data can be quantitative or qualitative. Secondary data used as basis for research are often quantitative and can be obtained from sources such as national statistical databases and government archives, and qualitative data can also similarly be found in some national data archives such as The Danish Data Archive.[22] When data are made accessible by depositing or sharing they can be reused for other research purposes, if primary data are reused they change into secondary. Below we elaborate further on the implications for the concepts of research object and research data.

4.2 Access to research data

A large international survey of researchers' practices and perceptions regarding data sharing finds that the majority of researchers from all subject disciplines have a positive attitude to sharing their own data and uses others'

(Tenopir et al., 2011). Only a minority, however, does so in the real world. Among the important conditions for agreement to share is a guarantee for authorship credit in the form of proper citation, offers of collaboration or of financial contributions to research. Among the reasons given for not making data available to others are lack of time, lack of funding, lack of places to put data and lack of standards. Still, at the same time, lack of access to other researchers' data is regarded as a barrier to scientific progress.[23]

A survey of researchers' practice at Aarhus University shows that they favour a more open access to research data and at the same time clearly express a need for better facilities for data storage and preservation (Sørensen et al., 2009). Similar results — in relation to issues of access — are found in an interview-based study of researchers at Glasgow University (Pattenden-Fail et al., 2010).

In a short summing up of the present circumstances, two trends are identified: at the same time as subject-specific data archives flourish, a lot of data lie on personal hard drives or are saved (or forgotten) elsewhere (Nelson, 2009). This is due to several factors: 'where', the lack of infrastructure, i.e., databases suitable for the various fields of science, 'how', basically the question of data standards, formats and metadata and 'which', raw data or quality controlled data?

An interpretation of these findings could be that issues of copyright in the sense of scientific recognition of authorship of data plays an important role, but we should also note the above-mentioned 'where', which points to insufficient data infrastructure and the 'how', which is the lack of common standards, both making data sharing cumbersome and time consuming for the researchers.[24] Whereas the issue of forms of recognition of authorship to research data is a matter that could be regarded as primarily belonging to the scientific community, the issues of infrastructure and data standards are rather a matter for information professionals in cooperation with the researchers.

But why share research data? The arguments for giving access to research data and thus facilitating sharing, use and reuse of data can be summarized thus, drawing on Borgman (2012):

1. Reproducing and verifying research results. This is, to some extent, an ideal standard, as not all experiments and field studies can be rep-

licated. In the sense of verification, however, it is stressed that scientific work is subject to tests of validity and reliability by the scientific community.
2. Serving the public interest. This can be seen both as the 'tax-payer argument' and as an argument for democratic transparency.
3. Enabling new questions to be asked for research data and results. This is related to 1, but not identical to it, as it emphasizes the possibility of reuse of data in the form of the combination of different data. It is also related to 2.
4. Advancing new research. The underlying argument is that data sharing can advance scholarship. This will, in turn, increase returns from investment in research. This is also related to 2.

Of vital importance for the access to and sharing of research data is that they can be discovered and retrieved. The condition for this is that they are described by means of metadata. As the aim of this study is not an in-depth discussion of metadata and their proper use, we limit ourselves to Higgins' typology of metadata (Higgins, 2012, p. 38):

- Descriptive metadata — ensures identification, retrieval, classification, links to related resources;
- Technical metadata — records file formats, software or hardware;
- Administrative metadata — information on acquisition, accession and issues of intellectual property rights;
- Use metadata — manages access, use statistics;
- Preservation metadata — documents preservation, migration.

4.3 Research data and research objects — the distinction revisited

It follows from the above reflections on data and objects that the original distinction between primary and secondary data cannot be maintained without modifications. If primary data are shared and reused by other researchers for other research aims, they must be regarded as secondary data. However, the aim could also be identical, but be for the purpose of the verification of earlier results (cf. 1.). The possibility of asking new questions to old data or of a reuse together with new (primary) data (cf. 3.) also blurs the original distinction between primary and secondary data.

Research objects, such as documents and recordings can be stored in digital archives and made searchable by issuing them with metadata. As we shall see later, this is the case for The State and University Library's digital archives such as LARM (LARM Audio Research Archive),[25] The Netarchive[26] and Mediestream.[27] In the case of LARM, users can already share annotations to sources (radio broadcasts) and in the case of Mediestream, users' enrichment of the available metadata through crowd sourcing is a future vision. We will later show that these cases illustrate that research objects can be issued with potential research data and, in turn, made accessible and shareable. This makes them at the same time research objects — and research data. It expands the scope of access to research data and brings the research libraries' and national archives' role into focus. We are also faced with the need for a wider understanding of the new nature of research data, facilitated by digital preservation. Indirectly, this also strengthens the argument for the use of the term 'research data' in the wide sense in the interviews of the DEFF-project.

5. E-infrastructure for the Danish national cultural heritage — targeted at researchers

Below we explore the ways in which the various digital preservation projects carried out by The State and University Library contribute to increasing the accessibility of research data. We concentrate on the services LARM, Netarchive, Mediestream, and DigHumLab. We will use the insight gathered from the DEFF project to explore the practice of data management with such question as: Are data preserved, curated, or 'only' stored? Which facilities are available?

Of special importance are the options for the users' adding of data to the digital objects, changing or enriching the metadata and sharing of data. These services provide access to original research objects (audio and video recordings, websites etc.) that are stored and preserved by The State and University Library. When they, for instance, are used and manipulated by researchers, enriched with additional data and eventually made accessible to other researchers, they change their nature and become both research objects and research data. These opportunities to change and add data requires that the storage system can handle the original data, as well as metadata about the changes and the manipulated data.

In our elaboration of the cases, we focus on the factual role of the library and on the types of institutions also involved in the work. We believe that it is necessary to consider whether a project is solely a project under the auspices of libraries, archives and museums (the LAM sector) or whether it also has participation from the research sector. Finally, we use our description of the services to demonstrate how a national and research library can play a vital role in research by establishing infrastructures suited to the researchers' needs.

The cases we present below illustrate how the library — here the State and University Library — collect the data while the researchers cooperate with the library in the development of software for the use of the data.

5.1 LARM.DK

LARM is a project active from 2010 to 2013.[28] In total, 11 different institutions are involved, among them 5 universities,[29] The Kolding School of Design, The Royal School of Library and Information Science, DeIC, the Danish national broadcast service (DR), The Museum of Media and The State and University Library. LARM has established a database with radio programs from 1925 to date. The purpose of LARM is to enable researchers to use radio programs as a source for humanistic research.

The State and University Library has been involved in LARM in different ways. One way has been as a source of data for the service, especially with scanned documents which has undergone OCR, in order to enrich the metadata of files originating from DR. By the end of 2013, the database will contain more than 1,000,000 hours of radio airtime. LARM has developed a platform to present the objects and data from the database, and to enable the researchers to annotate the metadata of a given program and to collaborate using each other's annotations. LARM is an example of research infrastructure, providing access to data and objects via metadata, with facilities for annotation, sharing and collaboration. Access to the database itself is limited to researchers. The annotations are shared as text under a Creative Commons License.[30] The archive itself is not intended as a facility for long-term preservation. However, the data are preserved by The State and University Library.

The platform which makes this possible is called CHAOS (Cultural Heritage Archive Open System[31]) and was developed by DR, The State and University

Library and the Danish Research Network.[32] The researchers access the radio programs via LARM.fm[33] when they are logged into the system; they can search the files, make their annotations and share these annotations. During this process the original objects are supplied with primary research data (or metadata), which, in turn, can be reused as secondary data by other researchers (Andersen, 2012a; Nordicom-Information, 2012).

LARM is an example demonstrating that close cooperation is needed between the LAM sector and the researchers in order to develop a system for access to digital research objects, to supply their metadata and share the enriched objects. The LAM sector has made both data and knowhow available regarding how to store large amounts of data and make them searchable. The LAM sector provided this project with knowhow regarding software programming.

5.2 The Netarchive — the Danish Web archive

Since 2005, The Royal Library and The State and University Library have harvested the Danish part of the dynamic internet.[34] From 2005 on, this has been required by law. The harvesting is done by using several mandatory strategies:

- Broad crawls: 4 times each year all websites from the domain.dk and Danish websites registered under other domains like .com, .org etc. are captured.
- Selective crawls: 80–100 specially selected sites are harvested. These are harvested with different intervals, some monthly, some up to six times a day.
- Event crawls: If a given event is rated as being of importance for the Danish society, and the pages are expected to disappear after the event, the archive harvests the related websites.

The Netarchive illustrates that institutions in the LAM sector are active with regard to important prerequisites for research infrastructures, in this case storage and long-term preservation of research objects which can be used by infrastructure services.

In 2012, 992,822 websites were harvested. This, of course, generated a huge amount of data. 88,542 GB are now stored in the archive. The material is issued with metadata and long-term preservation. The web pages thus harvested are archived as objects, which when used by researchers, are transformed into research data. Collaboration in the form of data sharing, as in LARM, is not yet possible. Only researchers can have online access to the Netarchive; if you are not a researcher but have a scientific purpose you can access it from the premises of the involved institutions. These limitations are necessary in order to avoid problems with copyright and to ensure the necessary protection of personal data (Andersen, 2012b; Schostag & Fønss-Jørgensen, 2012).

The software which gives access to the data is constantly being developed. At first, each researcher literally had to have personal assistance in order to search and use the archive. Then the Netarchive started to develop software, based on the Internet Archive Wayback Machine,[35] in order to improve online access. The archive is now involved in a project called the Digital Humanities Laboratory (DigHumLab),[36] whose goals, among others, are to improve the search facilities and to create new ways to analyze the materials in the archive. Below we shall go more into details about DigHumLab (Schostag & Fønss-Jørgensen, 2012).

The Netarchive illustrates that institutions in the LAM sector are actively working in fields that contain the key elements of infrastructures for the management of research data such as storage and long-term preservation of research objects and providing access for users.

5.3 Mediestream

Mediestream[37] is a project initiated in 2010 by The State and University Library. The project started as an attempt to digitize part of the national cultural heritage, which The State and University Library has in its collections. The project involves a newspaper collection, collections of TV and radio broadcasts and several other collections. Mediestream now contains TV programs that would take more than 87 years to watch. The newspaper collection contains 32 million pages from Danish newspapers from 1666 to today. The amount of data involved is massive, both in number of files involved and overall size. The newspaper collection will be digitized during the next couple of years.

The objects in Mediestream will be digitized material and material which was digitally born or later digitized as part of the archiving process. For example, the newspapers will be digitized using microfilms as the source. The TV programs will partly be analogue material digitized from VHS-based collections dating back from 1987, and partly be programs digitally harvested. Currently, Mediesteam offers access to Danish radio and TV programs broadcasted from 2006 to today, TV commercials from 1988 to 2005 and commercials shown in Danish cinemas from 1954–1995. The data will be long-term preserved.

For copyright reasons, the collection will be publicly accessible only on the premises of the Library, but it is online accessible for researchers and students (Elstrøm & Jensen, 2012; Kirring & Andersen, 2008; Williams, 2012).

In order to facilitate collaboration, The State and University Library would prefer it to be possible for researchers and students to enrich the metadata of the objects and share these annotations in both the Netarchive and Mediestream. This would transform the data for archival data into a dynamic database of research objects. Currently it is regarded as a technological problem, as a problem of quality with regard to the quality of the enriched metadata and as a copyright problem. In the future, LARM.fm could offer a solution to these.

Mediestream shows how national libraries can create collections of digital objects, which successfully can be made available for researchers and for the public as well. In this respect, Mediestream is primarily a facility for storage and preservation of research objects which, in turn, can provide content for research infrastructure services.

The three cases above illustrate the work processes involved when large collections of digital materials are made accessible. In order to facilitate searching and curating the data, librarians are needed to perform the traditional task of adding metadata. It is not just a question of adding them to the individual file or object, but also a question of combining different collections in order to create appropriate metadata. In other words, one has to use the knowledge of the librarians of the different collections libraries. Lauersen, Christiansen, and Olsen (2012), present how different sources are combined in order to enrich the metadata of files in LARM. In this case, it was possible to digitize printed material in order to create metadata so that LARM could present information on the programs broadcasted from as far back as 1925. The authors present a workflow which could make it possible to ensure that different collections are used in combination in order to produce metadata of as high a quality as possible.

5.4 Digital Humanities Laboratory (DigHumLab)

DigHumLab is a project with the aim to contribute to skills development, internationalization and innovation through a national focus on Digital Humanities in research, education and knowledge transfer. The aim is also to develop new methods and tools for analysis of data in the digital collections (Nordicom-Information, 2012).[38]

The project is carried out in cooperation between Aalborg University (AAU), Aarhus University (AU), The University of Copenhagen (KU), The University of Southern Denmark (SDU), the Royal Library and The State and University Library.[39] Besides these, the project also has international partners.[40]

The project addresses three different fields: Language Based Materials and Tools,[41] Media Tools[42] and finally Interaction and Design Studies.[43] We do not go into depth with each of these areas, but focus on Media Tools. This provides tools for two areas — each with a work package — which will enable the researchers to analyze web pages; also it will facilitate new ways of working with sound and media. Both Media Tools work packages involve The State and University Library. The first involves the Netarchive and the second involves the collections of films, TV and Radio that The State and University Library stores as part of its obligation as a national cultural heritage library (Nordicom-Information, 2012).

In order to ensure that the researchers involved in DigHumLab can access and use the data, The State and University Library is cooperating closely with the relevant researchers to develop the necessary tools. DigHumLab can be seen as an example of how the massive amounts of objects and data in the digital collections of national libraries best can be used by the researchers if they are involved in the development of tools for access and analysis.

5.5 Summary: access to research data and options for sharing and collaboration

The projects presented in this paper are summarized in Table 1.

Table 1: Overview of the projects.

Service	**LARM**	**The Netarchive**	**Mediestream**	**DigHumLab**
Purpose	To make Radio programs available as a source for research	To preserve Danish websites	To digitize and present cultural heritage online	To improve access to digital collections and new methods for analyzing the data
Content	Radio programs and other material from 1925 to today	Harvested websites	Different collections of The State and University Library	A wide range of online information, depending on the different projects
Participating institutions	Universities, institutions from the LAM sector, DR	The Royal Library and The State and University Library	The State and University Library	Universities, institutions from the LAM sector, DR
Data sources	DR, The State and University Library	The Netarchive	The collections of the participating institutions	Depending on the projects, among them the Netarchive and Mediestream
Access	Researchers — online	Researchers — online	Researchers — online and on the premises for others	Researchers — online
Long term storage	Yes	Yes	Yes	In the archives — yes
Options for users to annotate and collaborate	Yes	No	No	Part of the project is to ensure requirements for researchers' collaboration

The cases presented above demonstrate that a national library like The State and University Library, with a wide range of tasks, must be regarded as a natural part of a system which handles research data. In order to fulfil its tasks, the library has obtained experience, skills and knowledge regarding preservation of massive amounts of data, both in number of files and in size. These competences are created in cooperation between the LAM sector, researchers from the universities and DR, and are offered to other institutions outside the LAM sector via Digitalbevaring.dk.[44] The tradition of cooperation is continued in the FIF project mentioned earlier, in which DeIC and DEFF are evaluating systems and software to handle research data.

6. Conclusion and perspectives

The Danish universities clearly have a positive position on the issue of storage, preservation and sharing of research data. What for most universities prevents this from being translated into practice, is the lack of a national policy, diverse needs from the researchers' part depending on culture and traditions in the different fields of research, differences in IT-infrastructures and a general lack of financial incentives.

To some extent, the reuse of data as a consequence of data sharing blurs the classic distinction between primary and secondary data. In the case of digitally archived research objects such as broadcasts and websites, bringing together the examples of The State and University Library's services, the researchers' adding of metadata and annotations to the objects and making them shareable, adds new aspects to data sharing. This reflects a more collaborative research culture also inherent in the concepts of 'data deluge' and 'fourth paradigm for science' (the analysis of massive datasets) and also emphasizes the need for professional handling of copyright issues.

Providing storage and preservation facilities for researchers on an institutional and national level in the form of e-infrastructure(s) is the logical answer to the demand for a more open access to research data and research objects such as items of national cultural heritage. A viable e-infrastructure not only implies technology-based services for the description of data by metadata to ensure proper identification, facilitate retrieval and reuse, but also human expertise. The fields such as those of open access, copyright, metadata and archiving are evident. Less evident, but equally important, is the subject-specific knowledge of the different data collections. This is a prerequisite not only for the adequate description of data, but also for understanding the researchers' needs and the transformation of these into adequate technical solutions. This is best carried out in collaboration between the library staff and the researchers, as demonstrated by the cases from The State and University Library.

In general, the universities, the national libraries and national archives are responding to the challenge to improve the conditions for data management, data sharing and data preservation. DEFF, DeIC and several research libraries are involved in a project to evaluate different systems to handle research data. DeIC has been in contact with universities, national archives and research libraries in order to formulate a national policy regarding research data management. The universities themselves are formulating their own internal policies regarding research data. In all, one must expect that in a short time, the universities and the LAM sector will establish a coherent national system for the handling of research data.

References

Andersen, B. (2012a). Being a national library in a research infrastructure landscape. *Microform & Digitization Review, 41*(3–4), 175–179. doi:10.1515/mir-2012-0027.

Andersen, B. (2012b). *statistik_bja_2012.* Unpublished manuscript.

Borgman, C.L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059. doi:10.1002/asi.22634. Retrieved November 19, 2013, from http://onlinelibrary.wiley.com/doi/10.1002/asi.22634/pdf.

Christensen-Dalsgaard, B. (2013). *FIF – fælles infrastruktur for forskningsdata.* Unpublished manuscript.

DeIC. (2012). *Strategic goal: Coordinate solutions concerning data management and large data amounts.* Lyngby: Danish e-Infrastructure Cooperation. Retrieved November 19, 2013, from http://www.deic.dk/en/node/250.

DeIC. (2013). *Diskussionsoplæg – National strategi for datamanagement.* Lyngby: Danish e-Infrastructure Cooperation. Retrieved November 19, 2013, from http://www.deic.dk/DMdiskussionsopl%C3%A6g.

Donelly, M. (2012). Data management plans and planning. In G. Pryor (Ed.), *Managing research data* (pp. 83–104). London: Planet Publications.

Doorn, P., & Tjalsma, H. (2007). Introduction: Archiving research data. *Archival Science, 7*(1), 1–20. doi:10.1007/s10502-007-9054-6. Retrieved November 19, 2013, from http://link.springer.com/article/10.1007/s10502-007-9054-6/fulltext.html.

Elstrøm, G.V., & Jensen, T.S. (2012). Planning for mass digitisation of newspapers: A castle, a shed or something in between? *Microform & Digitization Review, 41*(3–4), 129–139. doi:10.1515/mir-2012-0021.

Hey, A.J.G., Tansley, S., & Tolle, K.M. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond, Wash.: Microsoft Research. Retrieved November 19, 2013, from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf.

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

Hox, J.J., & Boeije, H.R. (2005). Data collection, primary vs. secondary. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 593–599). New York: Elsevier. Retrieved November 19, 2013, from http://igitur-archive.library.uu.nl/fss/2007-1113-200953/hox_05_data%20collection,primary%20versus%20secondary.pdf.

Hyman, H.H. (1972). *Secondary analysis of sample surveys: Principles, procedures, and potentialities*. New York: John Wiley.

Kirring, E., & Andersen, B. (2008). Dansk radio og TV på statsbiblioteket. *DF Revy, 31*(3), 4–7. Retrieved November 19, 2013, from http://rauli.cbs.dk/index.php/revy/article/download/1733/1760.

Lauersen, D., Christiansen, K.F., & Olsen, L.L. (2012). Management of metadata for digital heritage collections. *Microform & Digitization Review, 41*(3–4), 151–158. doi:10.1515/mir-2012-0024.

Nelson, B. (2009). Data sharing: Empty archives. *Nature — LA English, 461*(7261), 160. doi:10.1038/461160a. Retrieved November 19, 2013, from http://www.nature.com/news/2009/090909/full/461160a.html.

Nordicom-Information. (2012). Aktuella forskningsprojekt. *Nordicom-Information, 34*(3–4), 105–136.

Pattenden-Fail, J., Sørensen, A.B., Kruse, F., Thøgersen, J., Molloy, L., & Ballaux, B. (2010). *Report on academic research practices*. London: Planet Publications. Retrieved November 26, from http://www.planets-project.eu/docs/reports/Reportonacademicresearchpractices.pdf.

Schostag, S., & Fønss-Jørgensen, E. (2012). Webarchiving: Legal deposit of internet in Denmark — A curatorial perspective. *Microform & Digitization Review, 41*(3–4), 110–120. doi:10.1515/mir-2012-0018.

Sørensen, A.B., Kruse, F., Thøgersen, J., Molloy, L., Pattenden-Fail, J., & Ballaux, B. (2009). *Report based on DT/7 questionnaire.* London: Planet Publications. Retrieved

November 26, from http://www.planets-project.eu/docs/reports/Planets_DT7-D4_Questionnaire_Report.pdf.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., et al. (2011). Data sharing by scientists: Practices and perceptions. *PloS One, 6*(6), e21101. doi:10.1371/journal.pone.0021101. Retrieved November 19, 2013, from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0021101.

Thestrup, J.B., Kruse, F., Nondal, L., Dorch, B.F., Andersen, M., Blaabjerg, N.J., et al. (2013). *Forvaltning af forskningsdata i danmark,* DEFF. Retrieved November 19, 2013, from http://udviklingspuljeprojekter.bibliotekogmedier.dk/sites/default/files/documents/Rapport_Forvaltning_af_forskningsdata.pdf.

Williams, K. (2012). Mediestream: The trials, tribulations and triumphs of making a digital collection available online. *Microform & Digitization Review, 41*(3–4), 171–174. doi:10.1515/mir-2012-0026.

Notes

[1] http://www.deff.dk/english/. Retrieved August 12, 2013.

[2] These topics are only parts of a data management plan, Donelly's checklist overview contains 10 items (Donelly, 2012, pp. 93–94).

[3] http://www.kb.dk/en/index.html. Retrieved August 12, 2013.

[4] http://www.cbs.dk/en/library. Retrieved August 12, 2013.

[5] http://library.au.dk/en/. Retrieved August 12, 2013.

[6] http://en.statsbiblioteket.dk/. Retrieved August 12, 2013.

[7] http://www.dtic.dtu.dk/English.aspx. Retrieved August 12, 2013.

[8] http://www.sdu.dk/en/Bibliotek. Retrieved August 12, 2013.

[9] http://www.en.aub.aau.dk/. Retrieved August 12, 2013.

[10] http://rub.ruc.dk/en/. Retrieved August 12, 2013.

[11] For a full list of the interview questions and sub-questions see Thestrup et al. (2013), Appendix 8.

[12] This does not include data that cannot be made public for data-protection reasons or data from private companies' research.

[13] See Doorn and Tjalsma (2007) for an overview of the development and recent trends of data archiving.

[14] http://rudar.ruc.dk, Retrieved August 12, 2013. https://data.kb.dk/dvn/, Retrieved August 12, 2013. http://www.larm-archive.org, Retrieved August 12, 2013. http://netarkivet.dk, Retrieved August 12, 2013.

[15] http://www.sim4rdm.eu/docs/project-outputs/sim4rdm-landscape-report. Retrieved November 1, 2013.

[16] http://www.deic.dk/node/110?language=en. Retrieved August 12, 2013.

[17] http://www.deic.dk/strategi. Retrieved August 12, 2013.

[18] http://dighumlab.dk/about/. Retrieved August 12, 2013.

[19] http://fivu.dk/en. Retrieved August 12, 2013.

[20] Part of this discussion is related to what is known as the "Fourth Paradigm of Science" (Hey, Tansley, & Tolle, 2009).

[21] We do not follow the axiom that primary data *per se* must be collected by the researcher(s) exclusively (e.g., Data Information Specialists Committee-UK, http://www.disc-uk.org/qanda.html) as this, in our opinion, tends to lead to a categorization based on copyright issues rather than on the relation to the research process (see also, Hox & Boeije, 2005).

[22] For an analysis of the emergence of national data archives see, e.g., Doorn and Tjalsma, 2007.

[23] We must remember that prestigious scientific journals such as American Economic Review, Nature and Science, for several years have required authors of articles to submit digital copies of documentation such as data, codes and protocols to the journal archives.

[24] This is also noted in the literature on the 'Fourth Paradigm' see, e.g., Hey et al. (2009).

[25] http://www.larm-archive.org/about-larm/. Retrieved August 12, 2013.

[26] http://netarkivet.dk/in-english/. Retrieved August 12, 2013.

[27] http://www.statsbiblioteket.dk/mediestream/. Retrieved August 12, 2013.

[28] http://www.larm-archive.org/about-larm/. Retrieved August 12, 2013.

[29] The University of Copenhagen, Roskilde University, The University of Southern Denmark, Aalborg University, Aarhus University, http://www.larm-archive.org/about-larm/. Retrieved August 12, 2013.

[30] http://creativecommons.org/. Retrieved August 12, 2013.

[31] http://www.chaos-community.org/en/index.html. Retrieved August 12, 2013.

[32] The Danish Research Network is today part of DeIC.

[33] http://www.larm-archive.org/about-larm/om-larm/om-larm-fm/. Retrieved August 12, 2013.

[34] http://netarkivet.dk/in-english/. Retrieved August 12, 2013.

[35] http://archive.org/web/web.php. Retrieved August 12, 2013.

[36] http://dighumlab.dk/. Retrieved August 12, 2013.

[37] http://www.statsbiblioteket.dk/mediestream/. Retrieved August 12, 2013.

[38] http://dighumlab.dk/about/mission-and-partners/. Retrieved August 12, 2013.

[39] http://dighumlab.dk/about/, Retrieved August 12, 2013 and http://dighumlab.dk/about/mission-and-partners/, Retrieved August 12, 2013.

[40] CLARIN (http://www.clarin.eu/), Retrieved August 12, 2013; DARIAH (http://www.dariah.eu/), Retrieved August 12, 2013; ESFRI (http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri), Retrieved August 12, 2013; TELEARC (http://www.noe-kaleidoscope.org/telearc/about/), Retrieved August 12, 2013 and IIPC (http://netpreserve.org/about-us/mission-goals), Retrieved August 12, 2013.

[41] http://dighumlab.dk/research-themes/language-based-materials-and-tools/. Retrieved August 12, 2013.

[42] http://dighumlab.dk/research-themes/media-tools/. Retrieved August 12, 2013.

[43] http://dighumlab.dk/research-themes/interaction-and-design-studies/. Retrieved August 12, 2013.

[44] http://digitalbevaring.dk/. Only in Danish. Retrieved August 12, 2013.