DE GRUYTER
SAUR

Filip Kruse,
Jesper Boserup Thestrup (Eds.)

RESEARCH DATA MANAGEMENT – AN EUROPEAN PERSPECTIVE

CURRENT TOPICS IN LIBRARY AND INFORMATION PRACTICE

DE
G

Based on case studies this book offers an insight into various European activities and practices in data management and their interaction with policies and programs. The latter form the background for the following case studies, provide the conceptual framework, at the same time giving an exhaustive understanding of the specific subjects. The case studies share common themes and give a concrete insight into vital issues such as web archiving, digitization of analog archives, researchers' motivations for sharing data, and how libraries, archives and researchers can collaborate in creating research tools and services.

THE SERIES: CURRENT TOPICS IN LIBRARY AND INFORMATION PRACTICE

This new series presents and discusses new and innovative approaches used by professionals in library and information practice worldwide. The authors are chosen to provide critical analysis of issues and to present solutions to selected challenges in libraries and related fields, including information management and industry, and education of information professionals. The book series strives to present practical solutions that can be applied in institutions worldwide. It thereby contributes significantly to improvements in the field.

www.degruyter.com
ISBN 978-3-11-036944-1
ISSN 2191-2742

CTLIP

Filip Kruse, Jesper Boserup Thestrup (Eds.)

RESEARCH DATA MANAGEMENT – A EUROPEAN PERSPE

DE G

Research Data Management

A European Perspective

Edited by
Filip Kruse and Jesper Boserup Thestrup

DE GRUYTER
SAUR

ISBN 978-3-11-036944-1
e-ISBN (PDF) 978-3-11-036563-4
e-ISBN (EPUB) 978-3-11-039606-5
ISSN 2191-2742

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Typesetting: RoyalStandard, Hong Kong
Printing and binding: CPI books GmbH, Leck
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

Current Topics in Library and Information Practice

Table of contents

Filip Kruse and Jesper Boserup Thestrup
Introduction — 1

Part I

Daniel Spichtinger and Jarkko Siren
1 **The Development of Research Data Management Policies in Horizon 2020 — 11**

Hans Jørgen Marker and Anne Sofie Fink
2 **CESSDA – a History of Research Data Management for Social Science Data — 25**

Veerle Van den Eynden
3 **What Motivates Researchers to Manage and Share Research Data — 43**

Part II

Hugo C. Huurdeman and Jaap Kamps
4 **A Collaborative Approach to Research Data Management in a Web Archive Context — 55**

Filip Kruse and Jesper Boserup Thestrup
5 **Archiving the Web – a Data Management Perspective — 79**

Piotr Filipkowski and Justyna Straczuk
6 **Revisiting the "lifestyle" Research and Creating a Qualitative Data Archive in Poland — 103**

Jonas Recker, Wolfgang Zenk-Möltgen, and Reiner Mauer
7 **Applications of Research Data Management at GESIS Data Archive for the Social Sciences — 119**

About the authors — 147
Index — 149

Filip Kruse and Jesper Boserup Thestrup

Introduction

1 About this book

This book presents selected cases related to research data management (RDM) in Europe.

Why Europe? In our opinion, much of the present data management discourse is dominated by contributions, presentations and examples largely based on American or British experiences. This is to no account unexplainable or unreasonable, given both the overall international impact of research from the two countries and the consequent more widespread tradition of data management. On the other hand, mainland Europe, with all its diversity in history, culture and academic traditions is in our opinion able to provide examples of RDM of no less significance, and perhaps also with unexpected elements.

Why present cases instead of for example a comprehensive cross-disciplinary analysis? A case is an event or other entity, situated in space and time and subject to empirical inquiry. This implies that context, content, intentions, actions, intended and unintended consequences are all components of the case. For the study of how RDM is carried out, it is in our opinion more important to understand the nuances of the various activities' reality, than getting an all-encompassing overview. Data management as activity is always connected with specific research activities or projects. Accordingly, knowledge of these is necessary for the understanding of data management, and as we shall see later, operations in data management weave in and out of the research process.

The cases here are not all tailored to the same pattern, but composed of recurring elements: the political and organizational framework, the researchers' roles and the outcome of the activities. The cases cover various aspects of RDM: the development of research tools for archives of the web, radio and television, social science data as well as the pioneering establishment of a national qualitative data archive, all closely linked to ongoing research projects. Other case studies present aspects of the development of research data policies, the evolution of social science data archives, the origins of web archiving and researchers' motivation to manage and share their data.

The chapters in the first part of the book provide the necessary framework for the subsequent chapters, as well as being case studies in their own right: the development of Horizon 2020; the history of CESSDA, the European organization of social science data archives, and of the interrelated evolution of data

DOI 10.1515/9783110365634-001

management in the social sciences; and finally an empirically founded analysis of the essential element in data management, namely the willingness of researchers to share data.

The chapters in the second part present case studies on web archiving practices in the Netherlands and Denmark, their history and the creation of research tools and services. This last mentioned activity in data management with special focus on the social sciences is studied exhaustively in the case of the German GESIS archives and their cooperation with researchers resulting in bringing social research to a new stage. The digitization and preservation of qualitative autobiographical social science data in the Polish Qualitative Data Archive offers an insight into the importance of analog sources in reconstructing social history.

2 A few remarks on data management

Practitioners of empirically founded research have always had to consider questions of which data to collect, how to analyze it and how to save the research results. To narrow down the historical perspective, the main dividing line runs through the meaning of "data". Today, data can be digital objects, typically in the form of audio, image or text files, but also various combinations of digital objects or structured information on these, in the form of datasets or databases, as well as digital information on physical objects. In this context, research data is any information in binary digital form derived from the research process.

The research process involves these basic activities and their corresponding phases: Research idea – planning of research – data acquisition – data analysis – publication of research results. This can be illustrated in various ways and with subdivisions of the activities or phases. Most often this is done in the form of a cycle[1].

The research idea or hypothesis is the research process in its embryonic form. When the research problem is formulated, it is at the same time an anticipation of decisions of methodology, data and data analysis, as well as of publication. The original idea is based on the researcher's knowledge of existing research results, her assessment of answered and unanswered questions, academic discourse in the field etc. This in turn guides the acquisition of data, be it in the form of collecting primary data, reuse of existing data or any combination thereof. Data analysis too, draws on experiences from earlier research results,

1 See eg. Pryor 2012; Ray 2014; Corti others et al. 2014.

as does the decision of selecting results for publication, deciding on publication form etc. Thus, data is at the core of every phase in the research cycle; and so is data management, albeit perhaps more concealed.

Moving away from these general considerations on research in favor of focusing on management of research data, various new aspects come into sight. Firstly, the growth in digital data, the "data deluge" (Pryor 2012); secondly the internationalization of research. This can be exemplified by the EU research and technological innovation programs FP7 (7th Framework Programme for Research and Technological Development) and its successor H2020 (Horizon 2020, The EU Framework Programme for Research and Innovation), with a total budget of 50 billion and 80 billion Euros, respectively. Both programs connect advances in science with technological and societal innovation leading to the third new characteristic, the increasing demand for social return of investments in science. This new utilitarianism is reflected both directly in the programmatic connection of science with societal interests and in the general condition for funding by H2020 that access to and re-use of research data and access to research publications should be made as open as by possible[2]. Further, data management planning is mandatory for certain projects and is in general strongly encouraged.

Two conclusions can be deduced from this development. The expansion of activities directly related to data extends the original basic research cycle with the phases of data preservation, data re-use and data sharing, with data curation as their common prerequisite. Data management planning becomes part of the research cycle, which now also can be viewed as a research data cycle, depending of the point of observation.

This development also encompasses data management planning as part of the two cycles. The creation of a data management plan is increasingly becoming a funder requirement but is also useful for the researcher, as it gives a systematic overview of how data will be handled during and after the research project. The overall elements of a data management plan can be summarized thus (Briney 2015, 20): characterization of data to be used in the project, documentation and organization of data, data storage during the research process, preservation of data after the project is completed and finally, future availability of data for sharing and re-use.

Part of data management planning is also the researchers' own clarification of their needs for curation of data, development of new tools, facilities for collaboration etc. The responsibility to meet these needs often lies with operators

2 http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide_en.pdf, accessed 06012017

external to the research process itself, such as data managers, data archivists and data librarians.

3 Contents of the book – an overview

The opening chapter of part one remedies the scarcity of research policy analyses, an odd scarcity as research is increasingly becoming the object of national, regional and international regulation in the form of funding for research programs in various fields, for academic exchange etc. Naturally, all these activities have strong connections with research policy on various levels.

Spichtinger and Siren give an in-depth analysis of how and through which processes the EU's research data policy developed. The 7th Framework Programme for Research and Technological Development (FP7) lacked a policy on research data, and the later report Riding the Wave[3] restricted itself to recommending the preparation of data management plans. In the preparation for Horizon 2020 (H2020) however, this played a pivotal part. An important factor in the creation of a policy was a European survey on scientific data, which proved the existence of widespread problems with access to research data, lack of funding for data infrastructures, and insufficient data policies on various levels. Furthermore, it showed strong support for making research data resulting from publicly funded research, freely accessible. Subsequently, it became obvious that open access to data needs to be balanced with issues of copyright, data security, data privacy etc. To put it in terms of a catchphrase: "As open as possible, as closed as necessary". Accordingly, a flexible pilot scheme for research data from EU funded projects, the Open Research Data Pilot (ORD)[4], was set up and recently extended to cover all areas of Horizon 2020. A key element of the Commission's open research data policy is a data management plan, which is mandatory for projects participating in the ORD pilot. This inclusion of RDM in EU research data policy is likely to promote proper data curation, storage, preservation, documentation etc. Initial experiences with data management plans in H2020 indicate that additional guidance for all actors in research projects, researchers, peer reviewers, funding administrators, data professionals etc. is necessary. Also, descriptions of data preservation, copyright and data standards frequently need to be developed.

3 https://ec.europa.eu/eurostat/cros/system/files/riding%20the%20wave.pdf, accessed 06012017
4 https://www.openaire.eu/h2020-oa-data-pilot, accessed 06012017

Providing access to and securing preservation of data is a task traditionally carried out by national data archives and is the basic prerequisite of all RDM activities as well as that of data based research, especially within the fields of social sciences and health.

Marker and Fink demonstrate that the social science data archives have evolved simultaneously with the expanding use of social science data in modern societies. RDM activities are thus from the beginning – albeit in embryonic forms – integrated in the archiving of social science data. The history of CESSDA (Consortium of European Social Science Data Archives), the umbrella organization for the European national data archives, indicates that already more than 40 years ago issues of data exchange, infrastructure and data were prominent on the agenda. It is important to note that at the time these activities did not attract political attention in the form of policy initiatives, but took place within the academic world of archival professionals. This can be regarded as a parallel to the development of EU research data policy as presented by Spichtinger and Siren, in the sense that the research world tends to lead its own life and that policy formulation in the field shows a tendency to be more reactive than proactive. CESSDA has since its pioneering days initiated and supported a wide range of projects aimed at software development, closer cooperation between the archives and the national statistical institutes with the aim of improving metadata standards and infrastructure. The vision for the future is to establish a seamless social science data archive service on a European level.

Making research data available for review, validation or reproduction of research results are all components of RDM[5]. While preservation of data can be regarded as the most fundamental prerequisite for these activities, this is hardly ever controversial for other reasons than those of funding and division of responsibilities. But what of researchers' own sharing of data? Are they willing to share? If so, for what reasons? Van den Eynden replies in the affirmative. Based on solid data she shows how data sharing benefits research, for example through enhanced visibility of the research and by ensuing prestige for its creator. Factors encouraging data sharing are cultural norms within the specific research community, external drivers such as funding and publishing policies and, perhaps surprisingly, the availability of data management services. Thus, an interesting potentially self-perpetuating mechanism is revealed. Not only do researchers' own data management activities such as sharing of data influence

5 A plethora of models of the research data life cycle exists. We take as our point of departure DCC's deliberative approach to the process of developing RDM services and of curating research data, see: http://www.dcc.ac.uk/resources/developing-rdm-services and http://www.dcc.ac.uk/resources/curation-lifecycle-model, accessed 06012017

the cultural norms and thus in turn the practice itself, but external institutional actors' provision of services for data management, training, support etc. stimulates data sharing, too. These findings are not limited to isolated fields of research but are valid across the wide range of academic fields.

Archiving the web and making it available for research presents challenges in itself related to its ephemeral nature, its content of personal data etc. These issues necessitate close cooperation between web archives, archivists and researchers.

The second part of the book opens with Huurdeman and Kamps' analysis of the interaction between existing web archiving practices in the Netherlands and the development of research tools[6]. The existing systems and interfaces for access facilitate primarily qualitative analysis, i.e. content analysis of single web pages, and not patterns and structures across multiple pages. The first step was to create a full-text search system instead of the existing system based on knowledge of the URL and date of the site or page. In addition, the researchers requested facilities for export of the data produced in order to use their usual tools and the adding of supplementary data (metadata) to enrich the original source. At the same time, development of functionalities within the web archive itself – although at the time not particularly in demand by researchers – also proved useful. The result of the collaborative efforts of researchers and web archivists were that both needs have been met, both for the big quantitative picture of large numbers of web pages and sites and for the smaller and deeper qualitative picture of individual items. In order to continue and accelerate this move from search engines to "research engines", Huurdeman and Kamp advocate increasing transparency and process support for researchers. Missing data in the web archive known to have existed by references in pages in the archive, must be located and included. This corresponds closely to the demand for increased transparency, also presented by the researchers. Support for the research process in the form of tools for corpus creation, analysis, dissemination and storage should preferably be provided within the system for access to the web archive, as this will improve data accessibility and potential reuse.

The theme of web archiving in cooperation with researchers is continued in Kruse and Thestrup's analysis of how new services for researchers in the web, newspapers, TV and radio have been co-created by researchers and the Royal Danish Library. Examples are the Royal Danish Library's DeIC Cultural Heritage

---

6 The analysis draws on experiences from several Dutch web research projects such as the webART project (2012-2016) which aimed at strengthening the accessibility of web archives. It was part of the larger CATCH program (2012-) with similar goals for the entire Dutch cultural heritage. See: https://www.nwo.nl/en/research-and-results/research-projects/i/07/7707.html and https://www.nwo.nl/en, accessed 06012017

Cluster and Mediestream. Facilities for full text search of the Danish web from 2005 onwards and creation of web spheres for search into metadata of broadcasts and into full text of digitized newspapers are now operational. Development of other features are in process or are already finished and implemented, such as integration and analysis of researchers' own data, sharing of data etc., while facilities for long-term storage remain outstanding. Behind this lies the internationally accepted maxim that all national media, regardless of format, are representatives of a nation's cultural heritage and thus rank on par with other artefacts. While the preservation and archiving of printed materials traditionally is managed by national libraries and regulated by laws on legal deposits, the archiving of other media has also become the task of national libraries. This, however, poses challenges related to e.g. the web material's ephemeral nature, general and specific legal restrictions in accessibility, technical issues etc. The analysis of the political process to meet these challenges shows both the seemingly permanent conflict between commercial interests, issues of copyright, issues of data protection and researchers' interests, as well as how the conflict was solved, at least temporarily.

It is a task of primary importance for social ethnographic research to save qualitative data documenting the lived life of ordinary people from oblivion, decay or destruction. Filipkowski and Straczuk relate the story of the Polish Qualitative Data Archive – part of the Polish Academy of Sciences – and analyze the scientific and political context of its genesis and further development. In spite of the rich Polish tradition of using autobiographical material in social research pioneered by Thomas and Znaniecki (1958), the continuation of the mission presented challenges. Much text and sound documentation has been lost, but recent attempts have succeeded in rescuing source materials and making them available for research. Of special interest is a set of field records on lifestyles of urban Polish families in the late 1970s. This study based itself on the interpretative sociological tradition and focused on the background of individuals and groups' motivations, choices and actions. Thus, it provides insight in the background for both the social and political situation of the time and of the changes later to come. Still, not every documentation from earlier studies can or should be preserved. Rather, while the Qualitative Data Archive still seeks to rescue data from oblivion or destruction, its aim is not to preserve every single piece of documentation with the risk of ending up with an overabundance of data, but to collect and curate selected sets of data and make them available for research. In short: the purpose is to build a reflective archive, an archive suitable for critical analysis of data.

Marker and Fink earlier established how social science data archives have evolved simultaneously with the use of data in modern societies and how RDM

activities from the beginning played an integrated role in the process. In continuation of this finding, Recker, Mauer and Zenk-Möltgen give an exemplary account of how the world of research and the world of archives are both nearing and supplementing each other. Their study shows how the GESIS Data Archive for the Social Sciences combines its role of being a "safe deposit box" for data with also providing development of tools for researchers' collaboration, data documentation, identification and publication. One example of the tools developed by GESIS is the datorium[7], a flexible online tool for the description and sharing of social science data. Several examples of RDM at GESIS are noteworthy. One is the European Values Study (EVS)[8]. GESIS is the official archive of this study and provides access to data and documentation of data. This requires that challenges of "metadata management" are met, that the software is viable and data documentation understandable, in order to enable re-use of data. Another example is the German Longitudinal Elections Study[9]. This project aims to provide researchers with fast access to well-documented data and transparent methodologies derived from various studies. To put it briefly, this chapter demonstrates that data archives are important partners in research as they are able to develop services that responds to the growing heterogeneity of research projects and their data.

References

Corti, Louise, Veerle Van den Eynden, Libby Bishop, and Matthew Woollard. 2014. *Managing and Sharing Research Data: A Guide to Good Practice*. Los Angeles: SAGE.

High level Expert Group on Scientific Data. 2010: Riding the wave. How Europe can gain from the rising tide of scientific data. A submission to the European Commission. https://ec.europa.eu/eurostat/cros/system/files/riding%20the%20wave.pdf accessed 06012017

Pryor, Graham. 2012. *Managing Research Data*. London: Facet.

Ray, Joyce M. 2014. *Research Data Management: Practical Strategies for Information Professionals*. Charleston Insights in Library, Archival, and Information Sciences. West Lafayette, Indiana: Purdue University Press.

Thomas, William I. and Florian Znaniecki. 1958. *The Polish Peasant in Europe and America: 2 Vol.* Unabridged and unaltered republication of the 2. ed. New York, London: Dover Publications

7 https://datorium.gesis.org/xmlui/, accessed 06012017

8 http://www.europeanvaluesstudy.eu/, accessed 06012017

9 http://www.gesis.org/wahlen/gles/, accessed 06012017