Bedankt voor het downloaden van dit artikel. De artikelen uit de (online)tijdschriften van Uitgeverij Boom zijn auteursrechtelijk beschermd. U kunt er natuurlijk uit citeren (voorzien van een bronvermelding) maar voor reproductie in welke vorm dan ook moet toestemming aan de uitgever worden gevraagd.

Behoudens de in of krachtens de Auteurswet van 1912 gestelde uitzonderingen mag niets uit deze uitgave worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch door fotokopieën, opnamen of enig andere manier, zonder voorafgaande schriftelijke toestemming van de uitgever. Voor zover het maken van kopieën uit deze uitgave is toegestaan op grond van artikelen 16h t/m 16m Auteurswet 1912 jo. Besluit van 27 november 2002, Stb 575, dient men de daarvoor wettelijk verschuldigde vergoeding te voldoen aan de Stichting Reprorecht te Hoofddorp (postbus 3060, 2130 KB, www.reprorecht.nl) of contact op te nemen met de uitgever voor het treffen van een rechtstreekse regeling in de zin van art. 16l, vijfde lid, Auteurswet 1912. Voor het overnemen van gedeelte(n) uit deze uitgave in bloemlezingen, readers en andere compilatiewerken (artikel 16, Auteurswet 1912) kan men zich wenden tot de Stichting PRO (Stichting Publicatie- en Reproductierechten, postbus 3060, 2130 KB Hoofddorp, www.cedar.nl/pro).

No part of this book may be reproduced in any way whatsoever without the written permission of the publisher.

info@boomamsterdam.nl
www.boomuitgeversamsterdam.nl
Data management planning and repository demands for qualitative research

René van Horik*

Research data management refers to the development, execution, and supervision of research plans, policies, programs, and practices that control, protect, deliver, and enhance the value of (research) data and information assets. The two main reasons to put effort into research data management are to protect the data from harm and to enable their reuse. Research data management, ideally based on a data management plan, looks into the specific issues of a research data set with respect to its long-term usability. Obviously the research data must be stored somewhere in order to enable their use and reuse. A trusted digital repository (TDR) is managed by an organization whose mission it is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future. Both concepts – research data management and trusted digital repositories – are covered in this article, with an emphasis on qualitative research data sets, containing interviews, image collections, text files, multimedia files, and the like. A good overview of issues related to the managing and sharing of research data is provided by Corti et al. (2014).

Archiving qualitative data sets

The practice of archiving social science research data sets emerged more than fifty years ago. Initially, quantitative research data sets were archived, and only in the mid-1990s were the first qualitative data sets archived. Slavnic (2013) describes the developments concerning data preservation and reuse of qualitative data in the UK and Sweden. The Dutch situation is described in “Fifty years of Dutch data archiving” (2014/2015).

The initial focus on quantitative research data was caused by the fact that computers in the ‘early days’ were number crunchers, rather than text processors. Nowadays, computer-assisted qualitative data analysis software (CAQDAS) assists the scholar in analyzing qualitative research data sets. Another reason for the late introduction of data archiving of qualitative research data sets lies in the complexity and diversity of qualitative data formats. This obstructed the emergence of standardized archive procedures. Legal issues, such as the protection of personal data in qualitative research data sets, also had an inhibiting effect on the usage of existing research data management infrastructures by qualitative researchers. Gebel et al. (2015) describe the situation in Germany concerning coping with priv-

* Dr. René van Horik works as program manager and researcher at DANS (Data Archiving & Networked Services) in The Hague. Email: rene.van.horik@dans.knaw.nl.
acy laws in social science research, in their article with the revealing title “Whatever is not explicitly permitted is prohibited”. Needless to say, the management of controlled access to qualitative data sets (e.g. by means of specific informed consent) is an important factor in removing the barriers for long-term access to the data. That the situation concerning the archiving of qualitative datasets in a standardized format is improving is illustrated by the fact that gradually more qualitative data sets are stored in trusted digital repositories. Currently over 25% of the about 2,000 oral history data sets stored in the trusted digital repository of Data Archiving and Networked Services (DANS) are available as open-access data sets in a standardized format. Access to most of the other data sets is granted after a user request has met certain conditions.

Data management planning

A research data management plan (DMP) is a formal document that outlines how researchers will handle data both during research and after the research is completed. The goal of a data management plan is to consider the many aspects of research data management, metadata generation, data preservation, and data analysis before the project begins. This ensures that data are well managed during the research process, and prepared for preservation, verification, and reuse in the future. Online tools are available to create a data management plan in an interactive way.

A data management plan grows out of an understanding of how data should be collected, normalized, processed, analyzed, preserved, used, and reused over their lifetime. Data management plans are often required by grant funding agencies. A data management plan that is associated with a research study must include comprehensive information about the expected data, such as the data types, the metadata standards used, the policies and facilities for access and data sharing, and the plans for data archiving and preserving so that it is accessible over time. Data management plans ensure that data will be properly documented and available for use by other researchers in the future.

Some specific data management issues related to qualitative data sets

A number of specific data management issues can be identified that require attention in relation to the archiving of qualitative data sets. It should be stressed, however, that they are not exclusively relevant for qualitative data. The three issues covered are (1) user license issues; (2) processing of data sets in order to enable reuse in the future; and (3) the coping with proprietary data formats created by CAQDAS software.

In principle the employer of the creator of a data set is the owner and copyright holder. If the creator uses existing qualitative sources, such as newspaper articles, photographs, or multimedia files, storage in a repository other than that in which the material is currently available is only possible after the owner of these sources
grants permission. Another license issue concerns the informed consent regulations that are part of any qualitative research project in which data of respondents is collected, for instance by means of interviews. The informed consent agreement between the researcher and respondents must cover whether, and under what conditions, the data will be accessible in the future. Quite often researchers promise respondents that they will delete data sets after the research is concluded in order to gain trust and increase the willingness of respondents to contribute to the research. Targeted deletion of data sets is indeed another aspect of research data management. However, academic norms regarding the verifiability of research results often impose a minimal data retention period of several years (although not necessarily as open data), and deleting data would make longitudinal studies impossible.

The second issue concerns the processing of data sets in order to facilitate their reuse. The anonymization of data sets is an example of this. By removing person names from the collected data, the privacy of the respondents is protected. Automatic ‘named entity extraction’ algorithms exist that can be used to anonymize data sets. The documentation of data sets is also an example of data processing. Several standards are available. Examples are the generic Dublin Core data element set aimed at online discovery of resources, and the DDI standard for the description of statistical and social science data.

The third issue is related to the usage of specific CAQDAS software. This is specialized software that is applied to structure and analyze qualitative data sets. Unlike quantitative research, which is based on numerical data with which calculations are performed and for which standardized data formats are available, qualitative research uses non-numerical qualitative data, such as texts, images, and multimedia. Qualitative data can be stored in a number of file types for which in many cases preferred and acceptable formats are available (such as TIFF for digital images or MP3 for audio files). CAQDAS applications are software programs that facilitate various research and analysis methods for the enrichment and analysis of qualitative data. Examples of CAQDAS applications are ATLAS.ti, HyperRESEARCH, MAXQDA, QDA Miner, NVivo, Qualrus, and Transana. Over the course of time, various releases of these applications have appeared, and there are versions for different operating systems (e.g. Windows, MacOS). Usually the files from these applications cannot be interchanged. CAQDAS applications use closed, proprietary database formats, and they do not support open import or export formats. This hampers the sustainability of the data files.

A few years ago, an open standard was developed to represent ‘richly encoded qualitative data’ and to serve as an archive format for CAQDAS files. The standard is known as QuDEx (Qualitative Data Exchange Format) and is available, inter alia, in the form of an XML schema. The exchangeability of qualitative data sets based on the QuDEx schema is described in Corti and Gregory (2011). Despite the fact that initially CAQDAS software vendors took interest in the QuDEx exchange scheme, support for the scheme as an exchange and archive format has not emerged until now.
Trusted Digital Repository

Several factors threaten the long-term usability of digital data. The data carrier (such as a CD-ROM or hard disk) can fail or become obsolete. Also, data formats (the way the data in the files is structured) risk becoming outdated, and this also holds for the software used to process the data files. Another threat concerns incomplete or a lack of documentation so one cannot assess the value and usability of a data set. Data sets can also get lost.

A trusted digital repository (TDR) has the mission to provide reliable, long-term access to managed digital resources to its so-called designated community, now and in the future. A designated community is an identified group of potential consumers who should be able to understand a particular set of information.

Concerning the long-term sustainability of research data, a number of criteria or guidelines can be distinguished (Data Seal of Approval, 2013). Trusted digital repositories have to comply with these guidelines. In the first place, the research data should be findable on the Internet by means of its documentation or metadata. Relevant legislation with regard to personal information and intellectual property of the data must be taken into consideration. The third criterion states that the research data must be available in a usable format. Next, the reliability of the data must be guaranteed over time. Finally, facilities for stable and robust references to the data sets (such as ‘persistent identifiers’) should be available.

A European Framework for Audit and Certification of Digital Repositories was set up to help organizations in obtaining appropriate certification as a trusted digital repository. It has established three increasingly demanding levels of assessment: Basic Certification, consisting of self-assessment and external review of the 16 criteria of the Data Seal of Approval (DSA); Extended Certification, including the Basic Certification and additionally an externally reviewed self-assessment against the more fine-grained ISO 16363 or DIN 31644 requirements; and Formal Certification, the validation of the self-assessment through a third-party official audit based on ISO 16363 or DIN 31644.

The re3data.org registry can be consulted in order to find a research data repository that facilitates the permanent storage of and access to research data in different academic disciplines.

Conclusion

Increasingly, funders and policymakers are requiring attention for research data management, for instance via the obligation to provide a data management plan as part of a research proposal. Services and infrastructures are available to assist researchers in the management of research data sets.

Privacy protection issues related to qualitative data sets, as well as the proprietary formats of databases managed by CAQDAS software, obstruct the long-term access and secondary analysis of qualitative data sets. Policies and procedures are available that can help to create a situation in which research data is stored in a trusted digital repository that enables long-term access according to current policies and
regulations with respect to the required consent of stakeholders, the need to anonymize data, and privacy laws that require the deletion of personal data. Nevertheless, situations will occur in which data sets have to be deleted because legal barriers obstruct their deposit in a repository. Determining that data storage is not feasible is also an aspect of research data management.

Concerning the long-term usability of proprietary formats (due to the fact that most CAQDAS software does not support open standards), data archives provide advice concerning the ingestion process of qualitative data sets and ‘preferred data formats’ that enable long-term access. The aspects mentioned above should be part of the formulation of a data management plan. An infrastructure of trusted digital repositories does exist in which data sets can be archived in an optimal way for the long term.

Notes

1. The description of terms and concepts related to research data is based on the glossary of Science Europe, edited by Peter Doorn and Rūta Petrauskaitė. See: http://sedataglossary.shoutwiki.com [Cited 2 November 2015].
2. See: http://www.dans.knaw.nl. The TDR of DANS can be found at http://easy.dans.knaw.nl [Cited 18 January 2016].
3. See: http://easy.dans.knaw.nl search for “Oral History”. 574 collections are open accessible. The access to 1,441 collections is restricted (situation as of 9 January 2016).
4. See for instance DMPOnline: https://dmponline.dcc.ac.uk/ [Cited 6 January 2016].
5. A template for a research data management plan can be found on the DANS website, see: http://dans.knaw.nl/en/about/organisation-and-policy/information-material [Cited 6 January 2016].
7. DDI, Data Documentation Initiative http://www.ddialliance.org/ [Cited 21 December 2015].
9. See: http://data-archive.ac.uk/create-manage/projects/quadex [Cited 21 December 2015]. The project so far has produced a proof of concept, including a QuDEx connection to the closed data base format of ATLAS.ti, but no practicable archival services have been developed based on the format. The standard is not yet supported by the manufacturers of the various CAQDAS application.
10. This concept is part of the “Reference model for an open archival information system” (ISO 14721:2012).
11. The 16 DSA guidelines can be found at http://datasealofapproval.org/en/information/guidelines/.
13. The registry can be found at http://service.re3data.org [Cited 5 January 2015].
References


