CentERdata
Institute for data collection and research

Data Management and Preservation Policy of the DNB Household Survey (DHS)

date	2 June 2020
authors	Marika de Bruijne, Maarten Streefkerk
version	3.0
classification	standard

Table of Contents

1 Introduction

This document outlines the data management and preservation policy for the DNB Household Survey (DHS). It explains the goals of the project and describes how data production, storage and dissemination functions are organized. Further, the document describes the measures taken to ensure the preservation of the project data for the long term.

2 Purpose

2.1 Mission

The DNB Household Survey (DHS) provides unique longitudinal data for the international academic community, with a focus on the psychological and economic aspects of financial behavior. The aim is to serve researchers by offering reliable long-term preservation and easy access to the data and metadata of the study.

2.2 Scope and Objectives

The DHS was launched in 1993 and comprises information on work, pensions, housing, mortgages, income, assets, loans, health, economic and psychological concepts, and personal characteristics. The data are primarily collected from 2,000 households participating in the CentERpanel, an Internet panel that reflects the composition of the Dutch-speaking population in the Netherlands.

The DHS data are made available online for all scientific researchers via the DHS Data Access website (see www.dhsdata.nl). The metadata are freely accessible and the data can be downloaded after registration. Use of the data is free of charge for scientific purposes.

In addition to using its own (meta)data dissemination system of the DHS, the data are archived in EASY, the online archiving system of the Dutch Data Archiving and Networked Services (DANS), to guarantee the long-term availability of the data.

3 Legal and Regulatory Framework

CentERdata, the owner of the DHS Data Access website, at all times complies with applicable laws and regulations including the Dutch GDPR (*Algemene Verordening Gegevensbescherming*). Furthermore, CentERdata uses working methods that meet the guidelines developed by the Association of Universities in the Netherlands (VSNU) as set out in the Code of Conduct for the use of personal data in scientific research (VSNU, 2005).

The panels to which the DHS study is administered are registered with the Dutch Data Protection Authority (*Autoriteit Persoonsgegevens*). The CentERpanel is registered under number m1274900. The LISS panel is registered under number m1332907. CentERdata is registered at the Tilburg Chamber of Commerce under number 41098659.

4 Organization

CentERdata, a research institute in the Netherlands, coordinates and implements the data collection of the DNB Household Survey (DHS). The Core Management Team of DHS governs the data collection, archiving and dissemination. An external Scientific Advisory Board (Wetenschappelijke Adviesraad) oversees and advises the CentERdata management team about the DHS.

CentERdata is a well-established organization internationally known for survey research. The financial support and active participation of the Dutch Central Bank (De Nederlandsche Bank, DNB) since 2003 emphasizes the DHS's societal importance, as does use of the DHS data by organizations such as the Netherlands Bureau for Economic Policy Analysis (CPB), the Netherlands Institute for Social Research (SCP), and the National Institute for Family Finance Information (NIBUD).

Several roles can be distinguished in the organization surrounding the DHS data (see Figure 1). In the following, we describe the roles and responsibilities according to three main functions within the data life-cycle: data production, data archiving & management, and data consumption (see also the illustration in Chapter 6, Figure 2). CentERdata both collects and archives the data of the DHS, which is why some of the roles can apply both to data production and to the data archiving & management tasks.

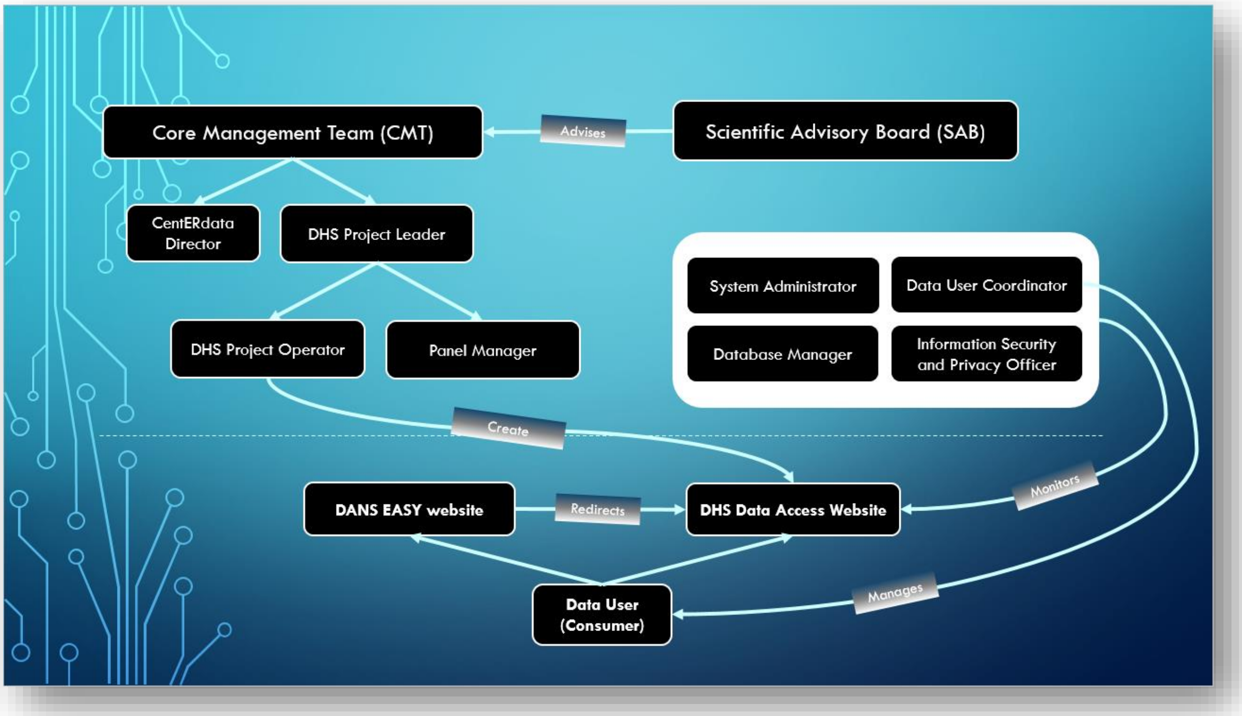

Figure 1. Organization of the DHS study

4.1 Data Production

The DHS is managed by a Core Management Team (CMT), headed by the director of CentERdata. The Scientific Advisory Board oversees the project and advises the Core Management Team.

Scientific Advisory Board (SAB)

The SAB advises on the design of the facility and provides feedback on newly planned investments, makes recommendations for changes and additions, and reviews both the scientific and societal contribution of the facility. The SAB meets at least once a year to discuss the content of the questionnaires and to respond to requests and suggestions for additions from the CentERdata Core Management Team. The SAB consists of Scientific researchers from DNB, academic experts from several universities and CentERdata researchers.

Core Management Team (CMT)

The DHS CMT consists of:

- CentERdata Director: The director of CentERdata makes the final decisions concerning the DHS study, but is obliged to solicit the SAB's advice. He/she will only deviate in exceptional cases and if doing so, will provide a motivation to the SAB. The director of CentERdata has the final responsibility for safeguarding the data.
- DHS Project Leader: The DHS Project Leader is responsible for the operational management of the study. He/she oversees the planning of data collection and ensures that all project members are familiar with and adhere to the data safeguarding plan. He/she reports to the director of CentERdata.

DHS Project Operator

The DHS Project Operator coordinates and implements tasks related to the DHS. He/she reports to the DHS Project Leader. For each Submission Information Package (SIP) there is a second-reader check before the SIP is accepted as an Archival Information Package (AIP, see section 4.2).

Panel Manager

A dedicated team manages the research panel(s), including support for and contact with the panel members (respondents). The panel manager coordinates tasks and employees within this team.

System Administrator

The System Administrator performs routine maintenance of the IT infrastructure and looks after the proper functioning of the servers.

4.2 Data Archiving & Management

DHS Project Leader

The DHS Project Leader is responsible for the data archiving and dissemination of the DHS data. He/she oversees the implementation of the archiving, data management and dissemination activities.

DHS Project Operator

The DHS Project Operator takes care of the operational data ingest activities and the dissemination of the metadata and data. He/she verifies the SIPs and converts them into Archival Information Packages (AIP). He/she accepts and publishes data updates on the DHS website. He/she also deposits the data disseminated via the DHS Data Access website into the EASY online archiving system of DANS. He/she monitors developments of new data formats and statistical tool versions and takes timely action to safeguard the long-term usability of the data and metadata.

Data User Coordinator

The Data User Coordinator receives and verifies the signed *CSS DHS data Statements* by Data Users (Consumers) and grants the Data Users access rights.

Database Manager

The Database Manager develops and maintains the dissemination system and online DHS Data Access web application. He/She also monitors developments in archival standards.

Information Security and Privacy Officer

The Information Security and Privacy Officer is responsible for the digital and physical security measures taken to ensure the safety and availability of the research data stored at CentERdata.

Partner: DANS

For an additional long-term preservation guarantee, the data disseminated via the DHS Data Access website are deposited in the DANS EASY online archiving system of DANS.

4.3 Data Consumption

Data User (Consumer)

Data Users (or Consumers) must agree to the user conditions set by CentERdata by signing the CSS DHS data Statement before being granted access to the data files.

5 Collaboration

Here, we briefly describe some of the main parties and collaborations involved with the DHS Data Access website.

5.1 DANS, DANS EASY, NARCIS

The data files and codebooks that are archived in and disseminated via the DHS Data Access website are also deposited in the EASY online archiving system of DANS (Data Archiving and Networked Services). Visitors of the DANS website have access to the metadata only when they either search NARCIS or the DANS EASY system. Search results lead to the DHS Data Access website for the actual data files, codebooks and more detailed metadata.

The metadata fields in the DANS EASY system are modelled as much as possible by the specifications of Qualified Dublin Core (see http://dublincore.org/documents/dcmi-terms/). Mandatory fields include: Title, Creator, Date created, Description, Access rights, Date available, Audience (the latter only in Standard).

5.2 Combell

To store the data, CentERdata uses the services of the hosting provider Combell. The data centers of Combell are fully ISO 27001 certified, redundant and physically located within the European economic area. CentERdata has its own protected partition, to which third parties have no access.

6 Data Process

This chapter explains the different tasks surrounding the DHS data storage and dissemination system, based on the OAIS (Open Archival Information System) functional model. According to the OAIS model, data processing can be divided into six functional entities and related interfaces (CCDS, 2012): 1. ingest, 2. data management, 3. archival storage, 4. access, 5. preservation planning and 6. administration (see Figure 2). In addition to these processes, we describe the pre-ingest phase which includes the data production.

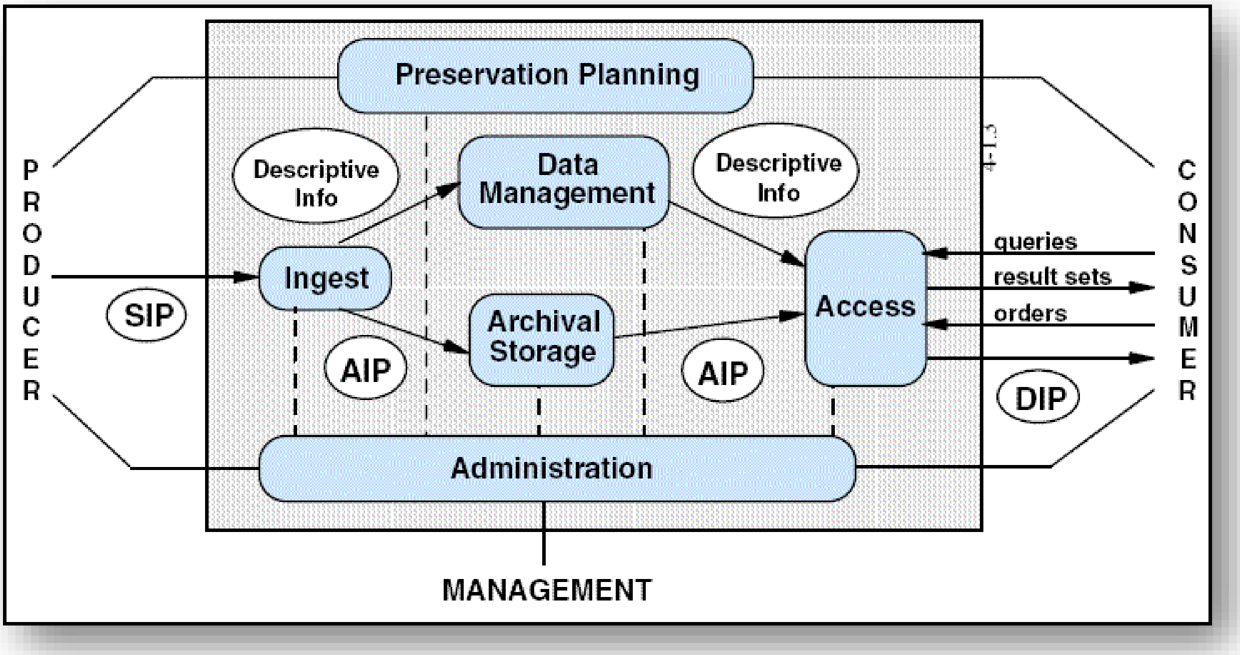

Figure 2. The OAIS model. NCDD (2013).

6.1 Data production

The purpose of the DNB Household Survey is to study the economic and psychological determinants of the financial behavior of households. The DNB Household Survey consists of five modules and is conducted every year. The topics included in the survey are work, pensions, housing, mortgages, income, assets, loans, health, and economic and psychological concepts.

Since 1993, CentERdata annually collects the DHS data through the CentERpanel, a web panel that consists of some 2,000 households. The CentERpanel is a probability-based household panel in the Netherlands. Households that could not otherwise participate are provided with a computer and Internet access.

Since 2017, the questionnaire is fielded among a sub selection of the LISS panel as well to keep up with demands of a high response. The LISS panel (Longitudinal Internet Studies for the Social sciences) consists of about 4,500 households, comprising almost 7,000 individuals. The panel is based on a true probability sample of households drawn from the population register by Statistics Netherlands (CBS). Households that could not otherwise participate are provided with a computer and Internet access.

As the DNB Household Survey is administered in the CentERpanel and the LISS panel of CentERdata, a formal consent procedure is required. The panel members consent via a multi-stage agreement including the initial recruitment as well as the activation of an account (after login only) in the panel environment. Since the introduction of the GDPR in 2018 panel members who already participated and newly recruited panel members, have been asked to give an explicit informed consent via a web form to (continue) taking part in research projects in the panel, among which the DNB Household Survey. Only respondents who complied could continue to participate in the panels.

The DHS is one of the few European micro-level panel datasets that permits detailed analysis of households' financial circumstances and economic behavior. In addition to providing a thorough picture of household wealth and debt portfolios in the Netherlands, the dataset is unique in that it allows researchers to identify important links between saving behavior and the psychological characteristics of individual household members.

Over 500 researchers around the world currently use data from the DHS in research projects and publications. A number of their papers have appeared in top-ranked journals such as the American Economic Review, Econometrica, Journal of Banking and Finance, Journal of Finance, and Journal of Financial Economics.

6.2 Ingestion

The DHS Project Leader is responsible for the correct data collection and processing of the raw data. After completion of the fieldwork, the data are labelled, cleaned from privacy sensitive information (e.g. open answers) and the respondents' answers are controlled for plausibility and outliers. This work is conducted by experienced survey researchers, accustomed to process and control survey data. The data processing steps are documented in (SPSS) syntax files/scripts. This ensures an audit-trail to the original data file. By running these scripts, the data are processed into a Submission Information Package (SIP) to be ingested by the DHS Data Access system.

In addition, the following procedures are formalized. For each SIP, there is an internal second-reader check. After this check, the SIP is entered into the repository and enriched with metadata. This Archival Information Package (AIP) is reviewed by a second reader. This second reader follows a Data Entry Checklist, which defines the required checks on the submitted data and metadata. Moreover, the data entry interface that is used to enter data into the DHS Data Access system, contains systematic checks to prevent the entry of duplicate data.

The data that are stored in and disseminated via the DHS Data Access system, are also deposited in the EASY online archiving system of DANS. Data Users have access to the metadata via the EASY system, but are redirected to the DHS Data Access website for the actual data files.

6.3 Archival Storage and System Architecture

CentERdata has developed its own system for the storage and dissemination of the DHS data, called DHS Data Access. This system is the technical basis of the DHS Data Access website and all surveys of the study are disseminated via this system. DHS Data Access is a web application built using a PHP framework that uses a relational database to store data. The DHS Data Access server is harvestable using an OAI-PMH implementation.

The public metadata of the studies in the DHS Data Access system support the main Dublin Core fields (Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language_id, Relation, Coverage, Rights).

6.4 Data Management and Administration

Within the context of the OAIS model, data management and administration concerns saving information on database requests and events as well as related statistical information. Also information on customer profiles and preservation process history is to be managed. This should enable tracking the migrations of AIPs, media replacements and AIP transformations.

In the DHS Data Access system, administrative information on database requests (data downloads) are logged and can be used to verify past events. To access the system, one must be uniquely logged in. External Data Users who are logged in gain limited rights to operate within the system, mainly to download the published datasets. Internally, CentERdata staff members also need to register to access the system. Depending on the tasks, a specific role is allocated to the staff member. The access rights within the system are dependent on this role.

Further, special attention is paid to two aspects of data management: ensuring data authenticity and version control. Data authenticity pertains to how the unchanged meaning and value of the data can be assured and verified.

When data files are created at the end of the data collection process, all data processing steps are documented in SPSS syntax files, which are stored in the same internal directory as the data files. Data file names include an extension which stands for the version number, and each time anything is altered in a data file it receives a new version number. The processing of new versions is also saved in syntax files. A brief description of the changes is given in a comment field next to the altered file on the website. This creates an audit-trail to the original data file and enables a reconstruction of the data processing if needed.

As part of the SIP, a metadata document referred to as a codebook is created. The version number of this document is embedded in the file name and given on the first page of the document.

In order to control the integrity of data files, of all uploaded files MD5 and SHA1 checksums are calculated when the file is uploaded to the server. It is possible to check the integrity of another copy of the data file by calculating the checksum of the data file and comparing its value with the checksum which was determined during upload of the published file. For example, Data Users can use the checksum to verify the integrity of the copy of the data file which they have downloaded in comparison with the version on the DHS Data Access website. The checksums are displayed alongside the files.

If the metadata or data need to be altered after ingesting the SIP into the data archive (as AIP), then the following procedure applies. The original SIP is modified by the DHS Project Operator. The DHS Project Leader then uses the same documentation procedure as for the first version, i.e. a syntax file is used for the data file including the modifications of the data file. A new version number is allocated to the file. After a check the Project Operator enters the new version of the file into the DHS Data Access system and enters information on the modifications into specified AIP fields which are visible for the Data Users. Old versions of data files remain stored in the database but are not visible to Data Users.

6.5 Access and Data Dissemination

Access to the DHS Data Access website is easy and open to every academic researcher, both in the Netherlands and abroad. The metadata and data are made available by CentERdata to scientific researchers through the website: https://www.dhsdata.nl

Metadata on each wave of the DHS study are freely accessible to the public on this website, including information on the study objectives, field work and metadata on the data file and individual variables. This information is included in the codebooks per wave.

While access to metadata is unrestricted, users must register in order to download actual data files. The Data User is required to sign and comply with the rules of the Statement concerning the use of CSS & DHS data, available at https://statements.centerdata.nl/css-dhs-data-statement.

The signed statement is verified by the Data User Coordinator, who sends the login information by email after access approval. The Data User can then download all published datasets within the database.

To enable meta crawlers to harvest the metadata of the DHS, the system supports the OAI-PMH protocol (base-url is https://www.dhsdata.nl/oai/). Dublin Core metadata information about published study units can be harvested here. The DHS metadata can also be searched by Google.

To increase visibility of the DHS data, the repository can be accessed through NARCIS, https://www.narcis.nl/ (“The gateway to scholarly information in the Netherlands”). Being the National Academic Research and Collaborations Information System, NARCIS is the main national portal for scientific information.

6.6 Long-term Preservation Strategy

An ISO 27001 certified hosting partner is responsible for the operational management of the server park and takes care of the tasks of the administration functional entity. The hosting partner also performs the updates of the software packages.

To address the threat of file format obsolescence, CentERdata monitors the development of the used software packages: SPSS and STATA, and also internally stores the data as *.csv files. In addition, the source code of the surveys, which includes the metadata to be used for labelling the data, is internally stored.

Besides in its own system, CentERdata archives the published data files and codebooks in the EASY system of DANS. The metadata deposited in the EASY system are defined on study level. While these data files are currently accessible to Data Users via the DHS Data Access website only, CentERdata has implemented a Statement of Intent with DANS to grant access via the EASY system, in case the DHS Data Access service should ever cease to exist. While the primary goal is to guarantee long-term preservation through good management of the DHS Data Access web application, this additional measure serves to create maximum trust in long-term preservation.

DANS creates persistent identifiers, URNs and DOIs, for the studies which are ingested by the EASY system. These can be viewed on the website of the EASY system. The URN of the DHS is also presented on the DHS Data Access website (see DHS Description).

7 Data Safeguarding

The DHS Data Access application is hosted on servers at the data center of Combell (see paragraph 5.2.). Access to the server rooms of Combell is limited by physical and organizational access measures, and the space is provided with fire protection and continuity guarantees for energy supply.

Functional access to the system is limited to the relevant system administrators. Concerning logical access security, the system is logically / organizationally protected by passwords. Account passwords are always hashed in the database after the first login. At CentERdata, access to the DHS Data Access application is based on role-specific authorization and only dedicated IT employees can access the underlying database of the DHS Data Access service. Concerning programmable security measures, the configurations and log files of the servers and applications used by CentERdata are periodically checked and updated where necessary. CentERdata keeps track of an incident registration.

An incremental backup of the DHS Data Access application (including data) is made daily to the back-up storage. The CentERdata application servers are protected by firewalls and measures have been taken to detect any irregularities on the network. The servers are also protected against DDOS attacks.

The security and risk management of the research databases of CentERdata, including the DHS Data Access system, is described in the CentERdata handbook on Information security and privacy. This document is based on the ISO standard NEN-ISO/IEC 27002 and is also in conformity with the Dutch “Code of conduct for use of personal data in scientific research”, published by the Association of Dutch Universities (VSNU). This handbook is available upon request.

8 Definitions

AIP

Archival Information Package. Submission Information Package (SIP) is ingested by the archive and processed into an Archival Information Package, which may contain more metadata than the SIP. An AIP conforms to the archive's data formatting and documentation standards (NCDD, 2012; CCSDS, 2012).

DIP

Dissemination Information Package. When a Data User requests information, the archive sends it to this information package which is derived from one or more AIPs (NCDD, 2012; CCSDS, 2012).

OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-threshold mechanism for repository interoperability (Open Archives Initiative, 2014).

OAIS

Open Archival Information System. An archive which has accepted the responsibility to preserve information and make it available for its designated community. The term 'Open' implies that the system-related recommendations and standards are developed in open forums, not that the access to the archive is unrestricted (CCSDS, 2012).

SIP

Submission Information Package. The data and the metadata which are sent by the Data Producer to the archive (NCDD, 2012; CCSDS, 2012).

9 References

CCSDS (2012). Reference Model for an Open Archival Information System (OAIS). Recommended practice, Issue 2. Washington, DC, USA.

DDI Alliance (2013). Website of the DDI Alliance. Information retrieved on 27 January 2014 from http://www.ddialliance.org/what.

NCDD (2012). Website Netherlands Coalition for Digital Preservation (NCDD). Information retrieved on 27 January 2014 from http://www.ncdd.nl/blog/?page_id=447.

Open Archives Initiative (2014). Website of the Open Archive Initiative. Information retrieved on 27 January 2014 from http://www.openarchives.org/pmh/.

VSNU (2005). Gedragscode voor gebruik van persoonsgegevens in wetenschappelijk onderzoek. Retrieved on 27 January 2014 from http://www.vsnu.nl/code-pers-gegevens.html.