CentERdata
Institute for data collection and research

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

date	15 July 2014
authors	Marika de Bruijne, Maarten Streefkerk, Lennard Kuijten, Eric Balster, Corrie Vis
version	1.1
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 an easy access to the data and metadata and reliable long-term preservation 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 collected from 2,000 households participating in the CentERpanel. The CentERpanel is an Internet panel that reflects the composition of the Dutch-speaking population in the Netherlands. Linked with other CentERpanel surveys, it generates substantial cost savings and rich research projects.

The data are made available online for all scientific researchers via the DHS Data Access website (see www.dhsdata.nl). The aim is to provide reliable and easily accessible information, including data and metadata, on the DHS. 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 Personal Data Protection Act (*Wet Bescherming Persoonsgegevens*). Furthermore, CentERdata uses working methods that meet the guidelines developed by the Association of Universities in the Netherlands (VSNU) as set out in Code of Conduct for the use of personal data in scientific research (VSNU, 2005).

The CentERpanel, to which the DHS study is administered, is registered with the Dutch Personal Data Protection Agency (*College Bescherming Persoonsgegevens*) under number m1274900. CentERdata is registered at the Tilburg Chamber of Commerce under number 41098659.

4 Organization

CentERdata, a research institute at Tilburg University in the Netherlands, coordinates and implements the data collection of the DNB Household Survey. The institute's management team governs over the data collection, archiving and dissemination, and CentERdata infrastructure has been tailored to facilitating these tasks. An external Scientific Advisory Board (Wetenschappelijk Adviesraad) oversees and advises the CentERdata management team about the DHS.

CentERdata is a well-established facility 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 such organizations 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 as well as the data archiving & management tasks.

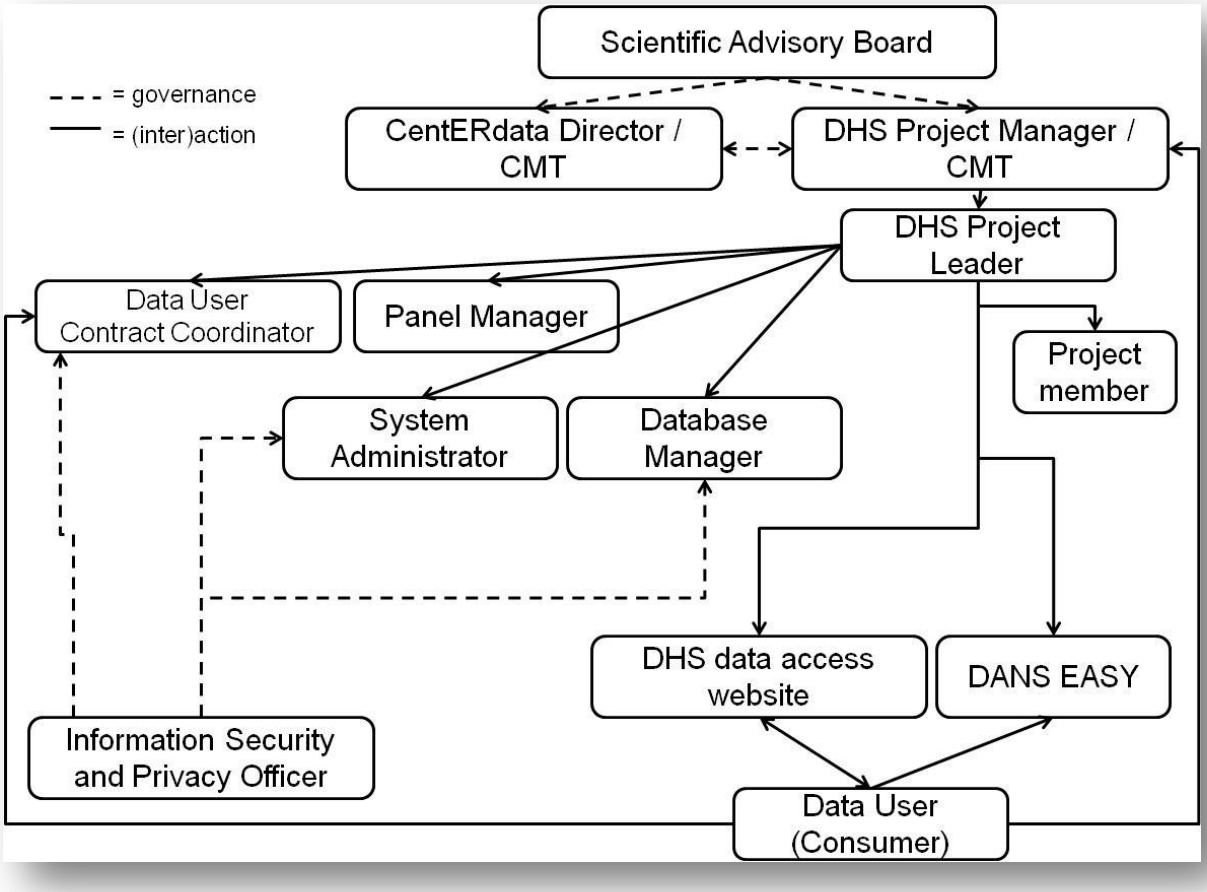

Figure 1. Organization of the DHS study

4.1 Data Production

Scientific Advisory Board / WAR

The SAB (Dutch: WAR) advises on the design of the facility but also 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 suggestions for additions from the CentERdata Core Management Team. The SAB consists of CentERdata employees and members external to CentERdata who have gained an academic expert position in the scientific fields related to the DHS.

Core Management Team

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. 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 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 Manager: The DHS Project Manager 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. The DHS Project Manager is also responsible for informing and requesting the consent of respondents and for maintaining the representativeness of the CentERpanel in which the data are collected. He/she reports to the director of CentERdata.

DHS Project Leader

The DHS Project Leader coordinates and implements tasks related to the DHS. He reports to the DHS Project Manager and is staff member of the Survey Research department. For each Submission Information Package (SIP) there is a second-reader check by another staff member of the Survey Research department before the SIP is accepted as an Archival Information Package (AIP, see section 4.2).

Panel Manager

A special department is dedicated to the operational management of the CentERpanel, including support for and contact with the panel members (respondents). The panel manager coordinates all tasks and employees within this department.

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 Manager

The DHS Project Manager 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. He/she is also responsible for the contracts with Data Users.

DHS Project Leader

The DHS Project Leader 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 coordinates the data-entry tasks of the Data Contract Coordinator. 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.

Data User Contract Coordinator

The Coordinator of the Data User Contracts receives and verifies the signed Contracts for the Use of Data 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 website. He also monitors developments in archival standards.

Information Security and Privacy Officer

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

Partner: DANS

For an additional long-term preservation guarantee, the data disseminated via the DHS Data Access website are deposited in the EASY online archiving system of DANS. An archive employee at DANS verifies the data and metadata which the DHS Project Leader has entered into their EASY system. If clarifications or corrections are needed, he contacts the DHS Project Leader before accepting the data into the system and publishing the metadata.

4.3 Data Consumption

Data User (Consumer)

Data Users (or Consumers) must agree to the rules set by CentERdata to regulate the appropriate use of the data by signing the Contract for the Use of Data before being granted access to the data.

5 Collaboration

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

5.1 DANS

The data that are archived in and disseminated via the DHS Data Access website are also deposited in the EASY online archiving system of Data Archiving and Networked Services (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, codebooks and more detailed metadata. One has to sign up before one can download datasets.

The metadata fields in the 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 VANCIS

A backup of database and web server files is made automatically every day and stored at VANCIS (formerly SARA), a Dutch super computer center. These data are stored on tape in a redundant manner and are divided over two different geographical locations. Recovery is only possible via a secured channel that only CentERdata has access to.

6 Data Process

This chapter gives 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): ingest, data management, archival storage, access, preservation planning and administration (see Figure 2). In addition to these processes, we describe here below the pre-ingest phase which includes the data production.

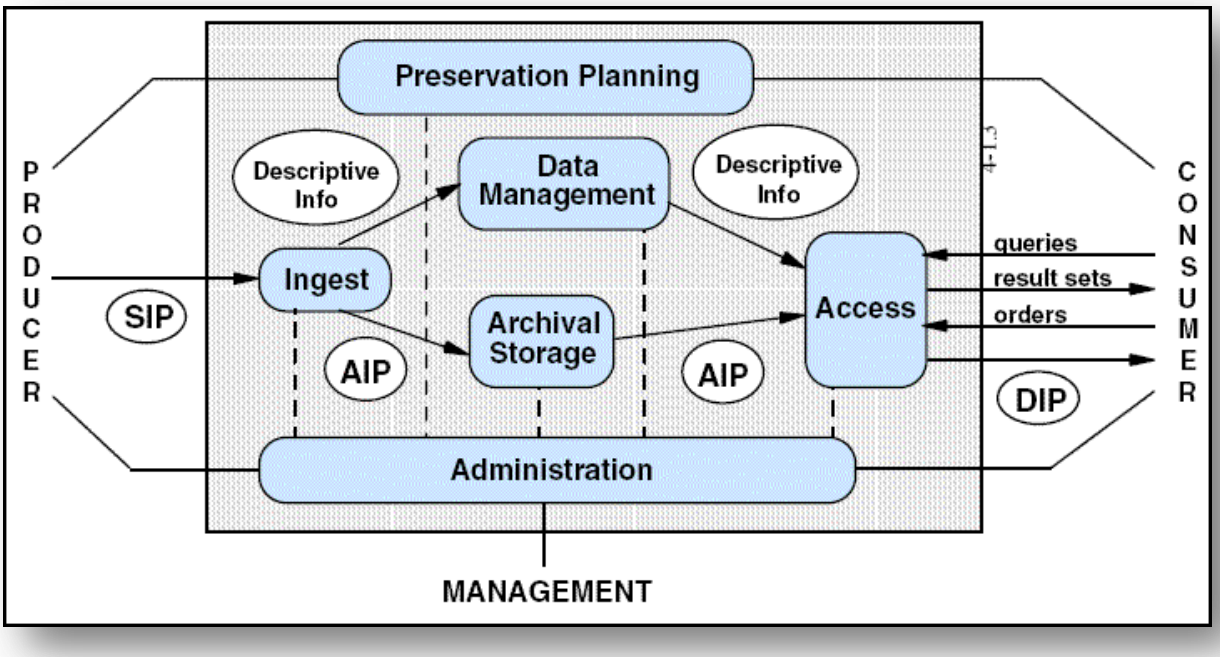

Figure 2. The OAIS model. NCDD (2013).

6.1 Data production

Since 1993, CentERdata annually collects economic data through a panel that consists of some 2,000 households, the CentERpanel. The purpose of this DNB Household Survey is to study the economic and psychological determinants of the saving 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.

The CentERpanel is a probability-based household panel in the Netherlands. Not all Dutch people have computers with Internet access, nor do the panel members. To ensure that the panel provides a good reflection of the Dutch-speaking population in the Netherlands, households without a computer and/or Internet access are given the use of a easy to use computer and Internet access. For more information about the CentERpanel, see: http://www.centerdata.nl/en/about-centerdata/what-we-do/data-collection/centerpanel

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 (1) it allows researchers to identify important links between

saving behavior and the psychological characteristics of individual household members, and (2) its Internet-based interface enables researchers to ask respondents topical questions with very short time lags. These short time lags are particularly attractive because the DHS is fielded in the CentERpanel. Members of this panel fill in questionnaires every week, and the detailed information in the DHS can easily be used in other research projects fielded in the same panel. The questionnaire design includes both questions on facts elicited on a recall basis (supported by available documentation) and hypothetical questions in experimental settings.

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 completing the field work, the DHS Project Leader processes the data into a Submission Information Package (SIP) to be ingested by the DHS Data Access system. All data processing steps are documented in and run using an SPSS syntax file to ensure an audit-trail to the original data file and a reconstruction of the data processing.

In addition, the following procedures are being formalized and implemented. To prepare the SIP, the Project Leader follows a procedure which is documented in the form of a checklist, containing data and metadata requirements and quality checks. For each SIP, there is an internal second-reader check. Before the SIP is converted into an AIP and accepted into the DHS Data Access system, the second reader follows a Data Entry Checklist, which defines the required checks on the submitted data and metadata. Moreover, the data-entry interface which is used to enter (meta)data into the DHS Data Access system contains systematic checks to prevent the entry of incorrect or duplicate (meta)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. These data are systematically entered into the EASY system by the DHS Project Leader. Once these data have been uploaded to the EASY system, a designated DANS employee verifies the data and if necessary will contact the DHS Project Leader, before the data are ingested into the EASY system. 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 supports the main Dublin Core fields (Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language_id, Relation, Coverage, Rights), used i.e. by the Data Documentation Initiative (DDI) standard. There are plans to migrate the DHS data under an application (Questasy) which is compatible with version 3 of the DDI.

Version 3 of the DDI introduces a life-cycle approach to documenting survey projects and distinguishes between the metadata of questions (data collection) and variables (dataset) (see DDI Alliance, 2013). For more information, visit the DDI Alliance website (http://www.ddialliance.org). CentERdata also collaborates with DANS to further develop data archiving and dissemination protocols. A more automatic data import from the DHS Data Access system into the EASY system is one of the project's aims for the coming future.

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 events and requests 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. Each data download of both external and internal users is logged and can be traced back to the individual user. Time stamps of any changes made to data and metadata are also logged.

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 (x.x), and each time anything is altered in a data file it receives a new version number. The changes between versions are saved in syntax files. A description of the changes is given on the website in a comment field next to the altered file.

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, alike to the protocol for data file names, as it is also given on the first page of the document. Changes between document versions are described at the beginning of each document.

In order to control the integrity of data files, of all uploaded files (data files, codebooks, images etc.) MD5 and SHA1 checksums are calculated as 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. Since the checksums are currently calculated by the system but not automatically displayed externally, the Data User can do this upon request. Internally, the checksums are also used to support version management.

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 Leader. Before starting to process the data file, the Project Leader compares the checksums of the published version and the copy which he/she will use for the new version. 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. If the content of a data variable needs to be changed, the variable receives a new name as the interpretation of the data variable might have changed. The changes to the content are documented in the related codebook, which is saved in the internal directory of the SIP. After a check the Project Leader 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.

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 are made available by CentERdata to scientific researchers through the website: http://www.dhsdata.nl, where the data files are also available.

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. The information is included in the beginning of the codebooks per wave.

While access to metadata is unrestricted, users must register in order to download actual data. The Data User is required to sign and comply with the rules of the Statement concerning the use of CSS & DHS data, available at http://www.centerdata.nl/sites/default/files/bestanden/dhs_statement.pdf

The signed statement is verified by the Data User Contract Coordinator, who sends the login information by e-mail 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 http://www.dhsdata.nl/oai/oai2.php). 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, http://www.narcis.nl ("The gateway to scholarly information in the Netherlands"). As the National Academic Research and Collaborations Information System, NARCIS is the main national portal for scientific information.

6.6 Long-term Preservation Strategy

The DHS strategy to reduce the risk of obsolescence is based on storing multiple copies on different storage media at different sites. If one of the sites collapses, this can be repaired by restoring the data from the other sites. To prevent sites from collapsing, all servers involved are placed in professional climate controlled server rooms.

Preservation ('planning functional entity') is secured further by backing up the data. All servers on which DHS data are stored are backed up daily. The backups are encrypted and stored at a different location. Since the data submitted to the DHS Data Access system is created by CentERdata, the Ingest functional entity is integrated in the systems of the archive. Its backup is made by VANCIS, a Dutch super computer center.

A System Administrator is responsible for the operational management of the server park and takes care for the tasks of the administration functional entity. The system administrator also performs the updates of the software packages.

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 website, this additional measure serves to create maximum trust in long-term preservation.

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

7 Data Safeguarding

7.1 Security and Risk Management

All data in the DHS Data Access system are stored on servers in an especially dedicated secured server room at Tilburg University. Only duly authorized Tilburg University server administrators and CentERdata server administrators have access to this room. To gain access to these servers, an administrator needs an electronic key and an alarm code, and must follow the procedure set out by the security officer of Tilburg University.

The security and risk management of the research databases of CentERdata, including the DHS Data Access system, are detailed 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.

7.2 Media Monitoring and Refreshing Strategy

All data in the DHS Data Access system are stored on redundant disk servers. These servers are monitored with a system that sends text messages to the system administrators on duty in case of a problem. As soon as a problem occurs, the system administrator can repair this using the redundant disk, or in case of a complete system crash, via the backup servers located at VANCIS.

The refreshing strategy consists of the periodical replacement of entire servers. These replacements are carried out based on the health and age of a server.

8 Definitions

AIP

Archival Information Package. Submission Information Package 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.