CentERdata
Institute for data collection and research

Preservation and Dissemination Policy of the LISS Data Archive

date	21 March 2016
authors	Marika de Bruijne, Arnaud Wijnant, Edwin de Vet, Eric Balster
version	1.3
classification	standard

Table of Contents

1 Introduction

This document outlines the data preservation and dissemination policy for the LISS Data Archive. It presents the purpose of the archive and the way in which the management tasks as well as the operational archival and dissemination functions are organized. Further, the document describes the measures which have been taken to ensure the security and preservation of the LISS data in the long-term.

2 Purpose

2.1 Mission

The LISS Data Archive preserves and disseminates the data which are collected in the LISS panel. The LISS panel was founded to facilitate research in the social sciences in the Netherlands and abroad. This is being done by providing scientific researchers with accessible data collection and data. The facility is open to academics anywhere in the world for scientific purposes.

2.2 Scope and Objectives

The data which are collected in the LISS panel are made available online for all scientific researchers via the LISS Data Archive (see https://www.lissdata.nl). The aim of the archive is to provide reliable and easily accessible information, including data and metadata, on the entire life-cycle of the LISS research projects.

In addition to its own archiving and (meta)data dissemination system, the LISS panel archives its data in EASY, the online archiving system of the Dutch Data Archiving and Networked Services (DANS), to ensure long-term availability of the data.

3 Legal and Regulatory Framework

CentERdata, the owner of the LISS Data Archive, in addition to the Netherlands Organization for Scientific Research (NWO), will at all times comply with applicable laws and regulations including the Dutch Data Protection Act (Wet Bescherming Persoonsgegevens). Furthermore, CentERdata uses working methods which are in accordance with the guidelines developed by the Association of Universities in the Netherlands (VSNU) in Code of Conduct for use of personal data in scientific research (VSNU, 2005).

The LISS panel is registered with the Dutch Data Protection Agency (het College Bescherming Persoonsgegevens) under No: m1332907. CentERdata is registered at the Tilburg Chamber of Commerce under the number KvK Tilburg: 41098659.

4 Organization

In 2006, the Netherlands Organization for Scientific Research (NWO) granted a proposal - An Advanced Multi-Disciplinary Facility for Measurement and Experimentation in the Social Sciences (MESS)- to set up the LISS panel (see NWO, 2006). In a later stage of the project, the Immigrant panel was set up as an additional panel, next to the LISS panel. In this document we will refer to both panels by the name LISS panel. The data collected in these panels are preserved and disseminated via the LISS Data Archive. The LISS Data Archive is managed by CentERdata, a research institute at the Tilburg University in the Netherlands.

Several roles are identified within the organization related to the LISS panel and data archive. Here below 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 1). CentERdata both collects and archives the data of the LISS Data Archive, which is why some of the roles can apply to both the data production as well as the data archiving & management tasks.

4.1 Data Production

Director CentERdata

> The director of CentERdata makes the strategic decisions concerning the panel and has the final responsibility for data safeguarding.

Head of Survey Research

> The Head of Survey Research at CentERdata is responsible for the operational management of the panel. He/she oversees the planning of data collection and makes sure all partners and data users are familiar and comply with the data safeguarding plan. The Head of Survey Research is also responsible for the informing and consent of respondents and for maintaining the representativity of the panel. He/she reports to the director of CentERdata.

Panel Manager

> A special department is dedicated to operational management of the panel, including support to and contact with the panel members. The panel manager coordinates these tasks and the employees within this department.

Project Leader

> Project Leaders in the data collection and dissemination projects of the LISS panel are usually provided by the Survey Research department. For each Submission Information Package (SIP) there is a second reader control by another employee of Survey Research before the SIP is delivered to the Data Archive Coordinator (see section 4.2). The Head of the Survey Research coordinates these employees.

System Administrator

> The system administrator performs routine maintenance of the IT infrastructure and guards the proper function of the servers.

4.2 Data Archiving & Management

Head of Survey Research

The Head of Survey Research is responsible for the data-archiving and dissemination of the LISS data. He/she oversees the implementation of the archiving, data management and dissemination activities. He/she is also responsible for the contracts with Client Researchers and Data Users.

Data Archive Coordinator

The Data Archive Coordinator takes care of the operational data ingest activities and dissemination of the metadata and data. He/she controls the SIPs and transforms them into Archival Information Packages (AIP). He/she coordinates the data-entry tasks of the Data Archive Employee. He/she accepts and publishes data updates on the lissdata.nl archive website and coordinates the depositing of the data disseminated via the LISS Data Archive into the online archiving system EASY of Data Archiving and Networked Services (DANS).

Data Archive Employee

The Data Archive Employee takes care of entering the data and metadata into the LISS Data Archive. He/she also carries out the data entry into the online archiving system EASY of Data Archiving and Networked Services (DANS).

Data Contract Coordinator

Data Contract Coordinator receives and controls the signed Contracts for the Use of Data for Data Users and grants the access rights to Data Users.

Database Manager

The Database Manager develops and maintains the archival system and the related online dissemination application. He/she also stays updated about the developments in the archival standards such as the DDI.

Information Security and Privacy Officer

The Information Security and Privacy Officer is responsible for the information and physical security measures that are taken to ensure the safety and availability of the archival data stored at CentERdata. He/she stays updated on the developments of new data formats and statistical tool versions to act on time in order to safeguard the long-term usability of the data and metadata.

Partner: DANS

For additional long-term preservation guarantee, the data disseminated via LISS Data Archive are deposited in the online archiving system EASY of Data Archiving and Networked Services (DANS). An archive employee at DANS controls the data and metadata which the CentERdata Data Archive Employee has entered into their EASY system. If clarifications or corrections are needed, he/she contacts the CentERdata Data Archive Employee before accepting the data into the system and publishing the metadata.

4.3 Data Consumption

Client Researcher

The Client Researcher gives an assignment to CentERdata to collect data in the LISS panel. Prior to data collection, he/she signs a contract with CentERdata on the data collection project and prior to receiving the data, the Statement for the Use of Data.

Data User (Consumer)

Data Users (or Consumers) comply with the rules stated by CentERdata on using the data in an appropriate manner by signing the Statement for the Use of Data of the LISS panel before being granted with the access to the data.

5 Co-operation

Here we briefly describe some of the main parties and co-operations which are related to the LISS Data Archive.

5.1 DANS

The data which are archived in and disseminated via the LISS Data Archive are also deposited in the online archiving system EASY of Data Archiving and Networked Services (DANS). Data Users have access to the metadata via the EASY system, but are referred to the LISS Data Archive for accessing the actual data files and more detailed metadata.

The metadata available via the EASY system are more limited than those available via the LISS Data Archive. While the LISS Data Archive contains metadata on question item and variable level, the EASY system contains metadata on a study level. The metadata fields in the EASY system are guided as much as possible by the specifications of Qualified Dublin Core (see http://dublincore.org/documents/dcmi-terms/). Obligatory 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 data center. Vancis BV is part of SURF, the collaborative ICT organization for higher education and research in the Netherlands. These data will be stored on tape in a redundant manner and are split over two different geographic locations. Recovery is only possible via a secured channel where only CentERdata has access to.

5.3 DDI Alliance

The DDI Alliance consists of many international organizations, including the national archives of several countries in North America, Europe, and Australia-Pacific. Questasy, the data dissemination tool developed by CentERdata and used for the archiving and dissemination of the LISS panel data, is based on the DDI 3 standard. To share her expertise and to follow the further developments within the standard, CentERdata collaborates with the DDI Alliance in several ways. This includes writing technical white papers, participating at expert workshops and collaborating with the DDI Technical Implementation Committee to improve the standard.

6 Data Process

In this chapter the different tasks around the LISS Data Archive are described applying the OAIS (Open Archival Information System) functional model. According to the OAIS model, the 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 1). In addition, we describe the pre-ingest processes which include the data collection.

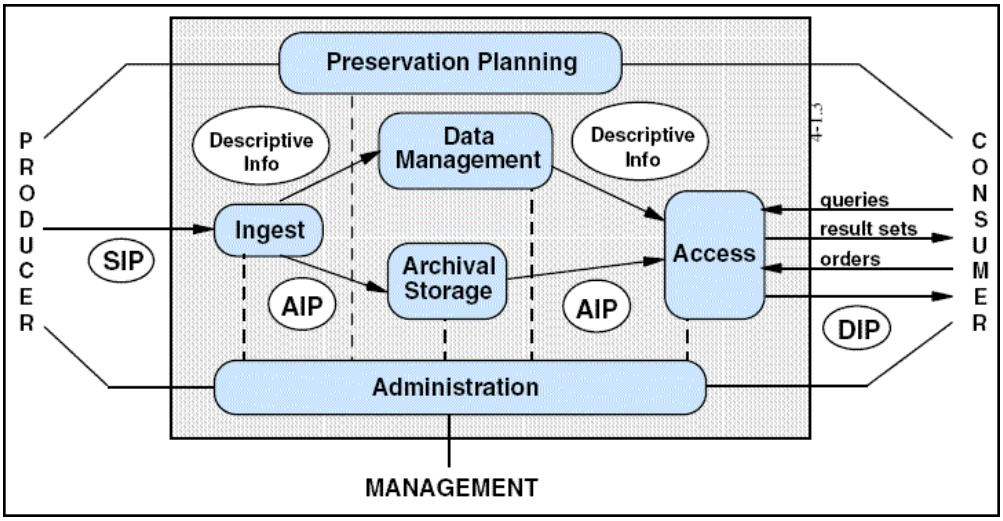

Figure 1. The OAIS model. NCDD (2013).

6.1 Pre-ingest

The data of the LISS Data Archive are collected in the LISS panel. The panel is an innovative data collection facility intended to boost and integrate research in various disciplines, such as economics, social sciences, life sciences, and behavioral sciences. The Longitudinal Internet Studies for the Social sciences (LISS) panel is a representative panel of about 4,500 households based on a probability sample drawn from population registers. Respondents answer interviews over the Internet monthly. Households that could not otherwise participate are lended a computer and broadband Internet access. Besides traditional questionnaires, the facility accommodates the use of visual displays, preloading of data, and the collection of non-interview data like self-administered measurement of biomarkers or ecological momentary assessment involving repeated sampling of subjects' current behaviors and experiences in real time, in subjects' natural environments.

Important elements of this infrastructure are its open access (to any academic researcher, both in the Netherlands and abroad) and its population-representativeness, providing an environment for cross-disciplinary studies and experiments on a wide array of topics and using advanced measurement devices.

Data Quality

Several studies on data quality have been carried out in the LISS panel. Comparing data from the CATI/CAPI recruitment interview for the LISS panel and data collected in the LISS online panel questionnaires, Scherpenzeel (2009) concluded that data collected via the Internet are at least as valid and reliable as those collected in the CATI and CAPI interviews. Validity coefficients estimated in a Multitrait-Multimethod (MTMM) model were higher for the Internet questions than for CAPI questions and similar to the CATI questions. Reliability coefficients obtained with the same model

were clearly higher for the Internet questions than for both the CATI and CAPI questions. Using a MTMM model as well, Révilla and Saris (2010) compared the quality of European Social Survey (ESS) questions in the (regular) face-to-face survey with the quality of the same questions in the LISS panel. Their conclusion was that the validity and reliability coefficients for ESS questions were similar in the Internet and the face-to-face mode of data collection.

Procedure

Academic researchers world-wide can request to collect data in the LISS panel. More detailed information on the procedure is available at: https://www.lissdata.nl/lissdata/how-use-panel

Once a research contract has been signed, a CentERdata employee will act as a Project Leader for the data collection project and will confer with the Client Researcher to coordinate the timing of the field work in the panel, as well as the questionnaire content and details. CentERdata has the right to revise or decline questions which it deems unsuitable for the panel members.

The CentERdata Project Leader is responsible for the correct data collection and processing of the data. After completing the fieldwork, the CentERdata Project Leader delivers the data to the Client Researcher. After this, he processes the data into a Submission Information Package (SIP) to be ingested by the LISS Data Archive.

6.2 Ingest

When a questionnaire has been fielded in the LISS panel, the concerned CentERdata Project Leader processes the raw data into a SIP. All data processing steps are documented in and run by using an SPSS syntax file to ensure a full audit-trail to the original data file and a reconstruction of the data processing.

For preparing the SIP, CentERdata Project Leaders follow a procedure which is documented in the form of a checklist, containing data and metadata requirements and quality checks. In addition, CentERdata is currently working on a general manual for data dissemination. For each SIP, there is an internal second-reader-control before the SIP is delivered to the Data Archive Coordinator.

When the Data Archive Coordinator receives the SIP, he controls the concerned data and metadata before ingest in the LISS Data Archive. Before the SIP is transformed into an AIP and accepted into the LISS Data Archive, the Data Archive Coordinator follows a Data Entry Check-list, which defines the controls to carry out on the data and metadata in submission. There are also several systematic checks within the data-entry forms used for entering (meta)data into the LISS Data Archive, which prevent entering incorrect or duplicate (meta)data.

The data which are archived in and disseminated via the LISS Data Archive are also deposited in the online archiving system EASY of Data Archiving and Networked Services (DANS). These data are systematically entered into the EASY system by a dedicated Data Archive Employee of CentERdata. When these data have been uploaded to the EASY system, a dedicated DANS employee controls the data and if necessary gets into contact with the CentERdata Data Archive Employee before the data are ingested into the DANS EASY system. Data Users have access to the metadata via the EASY system, but are referred to the LISS Data Archive for accessing the actual data files.

6.3 Archival Storage and System Architecture

CentERdata has developed its own system for data archiving and dissemination, called Questasy. This system is the technical basis of the LISS Data Archive and all questionnaires in the panel are disseminated via this system. Questasy is based on version 3 of the Data Documentation Initiative

(DDI). Version 3 of the DDI introduces a life-cycle approach of documenting survey projects and distinguishes between the metadata of questions (data collection) and variables (dataset). While earlier versions of DDI are widely used, no applications could be found in 2007 which applied version 3 to data as complex as the LISS data, leading CentERdata to build Questasy. See the DDI Alliance website for more information (http://www.ddialliance.org).

Questasy is a web application built using a PHP framework and uses a relational database to store data. The LISS Questasy server is also harvestable via an OAI-PMH implementation. Questasy source code is available to scientific, academic, and governmental non-profit organizations. More information on Questasy can be found at the following sources:
http://www.ddialliance.org/sites/default/files/QuestasyDocumentingAndDisseminatingLongitudinalDataUsingDDI3.pdf
http://www.iassistdata.org/downloads/iqvol3312amin.pdf
http://www.centerdata.nl/sites/default/files/bestanden/factsheet_ddi.pdf

Questasy has been developed to support a multi-panel support system, which is in accordance with the DDI-3 structure. A multi-panel support extension was necessary because of the addition of the Immigrant panel and anticipated the addition of other special groups or panels. The system also includes a so-called shopping basket, in which data users can put selections of variables from different LISS panel datasets and automatically obtain the merged data. The shopping basket function facilitates the exploitation of the longitudinal nature of the LISS panel data. In developing data dissemination protocols and procedures, close collaboration is aimed at with CESSDA (Council of European Social Science Data Archives) and DANS (Data Archiving & Networked Services).

6.4 Data Management and Administration

Within the context of the OAIS model, data management and administration include, among other, information on the database requests and events, statistical information needed by the archive administration and management, customer profile information and preservation process history information that tracks the migrations of AIPs, including media replacements and AIP transformations.

Concerning the administrative information on the database events and requests, these are logged by the application and can be used to verify past events. To access the data archival system, Questasy, one must be uniquely logged in. Data Users who are logged in, gain limited rights to operate within the system, mainly to download the published datasets and to view and edit parts of their personal account information.

Internally, CentERdata employees have to register for accessing the system and depending on the tasks, a specific role is allocated to the employee. The access rights within the system are dependent on this role. Each handling within the system is logged and can be trailed back to the individual user.

Further, special attention is paid to two aspects of data management: ensuring data authenticity and version control. Authenticity of the data concerns the means in which the unchanged meaning and value of the data can be ensured and verified. This is related to management of data versions and media monitoring.

The data deposited in the LISS Data Archive are data collected and managed within the LISS panel. 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 for each time anything is altered in a data file this receives a new version number. This procedure, including the file name which is saved, is included in the syntax files.

As part of the SIP, a metadata document, a codebook, is created. The file name of this document follows the same versioning procedure as that of the data file. Changes between document versions are described at the beginning of each document.

If the metadata or data need to be altered after ingesting the SIP into the data archive (as AIP), then the following procedure is followed. The original SIP is modified by the CentERdata Project Leader, using the same documentation procedure as for the first version, i.e. for the data file a syntax file is created including the modifications of the data file. A new version number is allocated to the file. When the description of a question item or variable label needs to be changed, this receives a new name, since the interpretation of the data variable might have been changed. The changes in the data are documented in a readme.txt document, which is saved in the same internal working directory as where the SIP is stored. After this the Project Leader delivers a new SIP version to the Data Archive Coordinator together with the readme.txt file. The Data Archive Coordinator enters the new version of the data into the data archive and enters information on the modifications, documented internally in the readme.txt file, into specified AIP fields which are visible for the Data Users. Old versions of data files remain stored in the database, but only the newest version of any file, such as the data file or codebook, is disseminated at a given moment.

To enable controlling on the integrity of data files, of all uploaded files (data files, codebooks, images etc), a MD5 and SHA1 checksum are calculated when the file is uploaded to the server. It is possible to check the integrity of the data file by recalculating the checksum of the current files on the server and compare those values with the checksum determined during upload of the file. This checksums are currently calculated by the system but by default not externally displayed. Upon request they can be provided to the Data User to view and control the integrity of the data file he has downloaded.

6.5 Access and Data Dissemination

Access to the LISS data is simple and open to every academic researcher, both in the Netherlands and abroad. In principle a few months after delivery to the original Client Researcher, the data are made available by CentERdata to scientific researchers through the LISS data website: https://www.lissdata.nl, where the LISS Core Study data are also available.

An extensive set of metadata on the whole life-cycle of the research project are accessible to the public on this website, including information on the study objectives, details on data collection, the entire questionnaire and metadata on the data file and individual variables. Also information on publications related to the data is provided when available. On the website one can use several ways to search within the database, such as free keyword search, browsing lists of studies, or a topic or concept based search.

While access to all metadata is unrestricted, it is necessary to register before being able to download actual data. The Data User needs to sign and comply with the rules of the Statement Concerning the Use of Data of the LISS Panel, available at https://www.lissdata.nl/lissdata/sites/default/files/afbeeldingen/LISS_statement_7.pdf

The signed statement is controlled by the Data Contract Coordinator, which sends the login information by e-mail at approval of access. The Data User can then download all published datasets within the database.

To enable harvesting the metadata of the LISS Data Archive, our repository supports the OAI-PMH protocol (base-url is https://lissdata.nl/oai/oai2.php). Dublin Core metadata information about published study units can be harvested here. The LISS Data Archive metadata can also be searched by Google.

To increase visibility of the LISS Data Archive studies, the repository is connected to NARCIS, http://www.narcis.info/ ("access to Dutch scientific information"), through which the LISS data can be found. NARCIS, National Academic Research and Collaborations Information System, is the main national portal for scientific information.

6.6 Preservation Planning and Long-term Preservation Strategy

The strategy to reduce the risk of files being unavailable is based on storing multiple copies on different storage media on different sites. Once 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 climate controlled professional server rooms.

Preservation (planning functional entity) is further secured by backing up the data. All servers on which LISS data are stored are backed up on a daily base. The backups are encrypted and stored on a different location. Because the data submitted to the LISS Data Archive is created by CentERdata, the Ingest functional entity is integrated in the systems of the archive. This backup is made by Vancis, the Dutch data center. Through this the archival storage function is implemented.

A System Administrator is responsible for the operations of the server park. This includes the tasks of the administration functional entity. Also the updates of the software packages are done by this system administrator.

In addition to its own system, CentERdata archives the published data files and codebooks in the EASY system of DANS, including study level metadata as is contained in the EASY system. While these data files are currently accessible for Data Users only via the LISS Data Archive, CentERdata had signed an agreement of intention with DANS to grant access via the EASY system in case the existence of the LISS Data Archive would ever become jeopardized. While the primary goal is to guarantee long-term preservation by good management of the LISS Data Archive, this additional measure aims at creating maximum trust in long-term preservation.

Currently, DANS creates persistent identifiers, in this case URNs, for the LISS data files when they are ingested by the EASY system. These can be viewed on the website of the EASY system. CentERdata has the intention to implement URNs also in its own system. CentERdata is working together with DANS for a solution on how to implement persistent identifiers in the future.

7 Data Safeguarding

7.1 Security and Risk Management

All data in the LISS Data Archive are stored on servers in a specially dedicated secured server room at the Tilburg University. Only the Tilburg University server administrators and CentERdata server administrators who are authorized for the task have access to this room. To get access to these servers, an administrator needs an electronic key, alarm code and should follow the procedure from the security officer of Tilburg University.

The security and risk management of the LISS data archive is covered by the CentERdata handbook Information security and privacy. This document is based on the ISO standard NEN-ISO/IEC 27002 and is also in concordance 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 within the LISS data archive are stored on redundant disk servers. These servers are monitored with a system that sends text-messages to the system administrators in duty in case of a problem. Once a problem occurred, 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 complete servers. This replacement is considered 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, 2013; CCSDS, 2012)

DDI

The Data Documentation Initiative (DDI) is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. The DDI metadata specification supports the entire research data life-cycle. (DDI Alliance, 2013)

DIP

Dissemination Information Package. When a Data User requests information, the archive sends this to this information package which is derived from one or more AIPs. (NCDD, 2013; CCSDS, 2012)

OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability (Open Archives Initiative, 2013).

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, 2013; 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 5 April 2013 from http://www.ddialliance.org/what.

NWO (2006). An Advanced Multi-Disciplinary Facility for Measurement and Experimentation in the Social Sciences (MESS). Retrieved on 20 May 2013 from http://www.nwo.nl/onderzoek-en-resultaten/onderzoeksprojecten/80/2300133180.html.

NCDD (2013). Website Netherlands Coalition for Digital Preservation (NCDD). Information retrieved on 5 April 2013 from http://www.ncdd.nl/blog/?page_id=447.

Open Archives Initiative (2013). Website of the Open Archive Initiative. Information retrieved on 5 April 2013 from http://www.openarchives.org/pmh/.

Révilla, M.A. & Saris, W.E. (2010). Comparison of surveys using different modes of data collection: European Social Survey versus LISS Panel. Working paper, New Developments in Survey Methodology – seminar series. Universitat Pompeu Fabra, Spain.

Scherpenzeel, A.C. (2009). Online interviews and data quality: A multitrait-multimethod study. Working paper, CentERdata, Tilburg University.

VSNU (2005). Gedragscode voor gebruik van persoonsgegevens in wetenschappelijk onderzoek. Retrieved on 20 May 2013 from http://www.vsnu.nl/code-pers-gegevens.html.