Preservation and Dissemination Policy of the PROFILES Registry

date	1 March 2016
authors	Nicole Horevoorts
version	1.1
classification	standard

Table of Contents

1 Introduction

This document outlines the data preservation and dissemination policy for the PROFILES registry. It explains the purpose of the archive and describes how management tasks and the operational archival and dissemination functions are organized. Further, the document describes the measures taken to ensure the security and preservation of the PROFILES registry for the long term.

2 Purpose

2.1 Mission

The PROFILES registry preserves and disseminates the data that are collected in the PROFIEL studies. The PROFIEL studies are done with help of the PROFIEL data manager tool, which was developed to facilitate research in the social sciences in the Netherlands and abroad. This is done by providing scientific researchers with accessible data collection and data. The PROFILES registry is open to academics anywhere in the world for scientific purposes, free of charge. It is a public good, with maximum scientific and societal payoff.

2.2 Scope and Objectives

The data that are collected in the PROFIEL studies are made available online for all scientific researchers via the PROFILES registry archive (see http://www.profilesregistry.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 PROFIEL studies.

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

3 Legal and Regulatory Framework

Tilburg University and the Comprehensive Cancer Centre the Netherlands, the owners of the PROFILES registry, jointly with the Netherlands Organization for Scientific Research (NWO), at all times comply with applicable laws and regulations including the Dutch Personal Data Protection Act (*Wet Bescherming Persoonsgegevens*). Furthermore, they use 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 PROFIEL study is registered with the Dutch Authority Personal Data (*Autoriteit Persoonsgegevens*) under number m1433221.

4 Organization

In 2008, the Netherlands Organization for Scientific Research (NWO) granted a proposal - Building an infrastructure for multidisciplinary and longitudinal data collection of the physical and psychosocial impact of cancer and its treatment: Patient Reported Outcomes Following Initial treatment and Long term Survivorship (PROFILES) - (see NWO, 2008). The data collected in through this infrastructure - the PROFIEL studies - are preserved and disseminated via the PROFILES Registry Data Archive. The PROFILES Registry Data Archive is managed by the Comprehensive Cancer Center the Netherlands and Tilburg University in the Netherlands.

Several roles can be distinguished in the organization surrounding the PROFIEL studies and the PROFILES Registry data archive (hereby named PROFILES, unless specification is needed). 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). PROFILES both collects and archives the data of the PROFILES Registry 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

Executive Team (ET)

PROFILES is managed by an Executive Team. An external advisory board oversees the project and advises the ET: the Advisory Board.. The ET coordinates the project and communicates and confers regularly with the Advisory Board.

Advisory Board (AB)

The AB 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 AB performs these tasks by annually receiving an update in a meeting. In this meeting future plans and developments are also discussed. The goal is to have a large number of disciplines represented in the boards, with members that are international experts in the field.

Directors

Because of the dual ownership of PROFILES there are two directors: the director of the Medical and Clinical department of Tilburg University and the director of the Comprehensive Cancer Center the Netherlands. These directors have the formal responsibility of PROFILES. They will only deviate from ET's decisions in exceptional cases and if doing so, the will provide a motivation to the ET and AB. The directors have the final financial responsibility.

PROFILES Project Manager

The PROFILES Project Manager is responsible for the operational management of the panel. He oversees the planning of data collection and ensures that all partners and data users are familiar with and adhere to the data safeguarding plan. The PROFILES Project Manager is also responsible for informing and requesting the consent of respondents and for maintaining the representativeness of the panel.

PROFILES Project Researcher

The PROFILES Project Researcher coordinates tasks and projects relating to the PROFILES panel. He reports to the PROFILES Project Manager.

Panel Manager

A special department is dedicated to the operational management of the panel, 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. For PROFILES this role is performed by CentERdata.

4.2 Data Archiving & Management

PROFILES Dissemination Manager

The PROFILES Dissemination Manager is responsible for the data archiving and dissemination of the PROFILES 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 and grants the Data Users access rights.

Data Archive Employee

The Data Archive Employee takes care of the operational data ingest activities and the dissemination of the metadata and data. The Data Archive Employee enters the data and metadata into the PROFILES Registry Archive and publishes data updates on the profilesregistry.nl archive website. He/she also coordinates the depositing of the data disseminated via PROFILES Registry Archive into the EASY online archiving system of DANS.

Database Manager

The Database Manager develops and maintains the archival system and the related online dissemination application. He/she also is responsible for the information and physical security measures taken to ensure the safety and availability of the archival data stored at PROFILES. 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 .For PROFILES this role is performed by CentERdata.

Partner: DANS

For an additional long-term preservation guarantee, the data disseminated via the PROFILES Registry Data Archive are deposited in the EASY online archiving system of DANS. An archive employee at DANS verifies the data and metadata which the PROFILES Registry Data Archive Employee has entered into their EASY system. If clarifications or corrections are needed, he contacts the PROFILES Registry 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 PROFILES to collect data in the PROFIEL studies panel. Prior to data collection, he signs and undertakes to comply with the rules in the agreement that is set up for the Client Researcher.

Data User (Consumer)

Data Users (or Consumers) must agree to the rules set by PROFILES to regulate the appropriate use of the data by signing the Statement Concerning 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 LISS Data Archive.

5.1 CentERdata

CentERdata has developed the PROFIEL data manager tool and the PROFILES data dissemination tool. CentERdata also manages the databases on which the PROFIEL studies data are saved.
CentERdata has received the Data Seal of Approval in 2010 through their LISS Panel. The method of PROFILES is based on the LISS Panel method.

5.2 DANS

The data that are archived in and disseminated via the PROFILES Registry data archive 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 referred to the PROFILES Registry data archive for the actual data files and more detailed metadata.

The metadata available via the EASY system are more limited than those available via the PROFILES Registry data archive. While the PROFILES Registry data archive contains metadata at question item and variable level, the EASY system contains metadata at the study level. The metadata fields in the EASY system are modeled 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).

5.3 VANCIS

A backup of database and web server files is made automatically every day and stored at VANCIS (formerly SARA), the 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.

5.4 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 PROFILES Registry data archive, is based on the DDI 3 standard. To share its expertise and to monitor 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

This chapter describes the different tasks surrounding the PROFILES Data Archive, applying 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 1). In addition, we describe the pre-ingest processes which include the data collection.

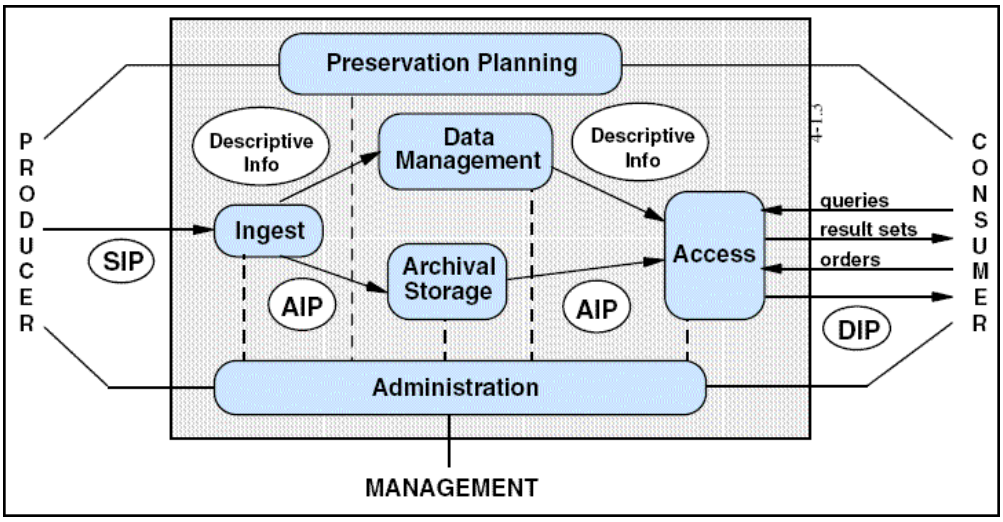

Figure 1. The OAIS model. NCDD (2013).

6.1 Pre-ingestion

The data of the PROFILES Data Archive are collected in the PROFIEL studies panel. The PROFIEL studies panel currently contains more than 15,000 cancer survivors of whom we received more than 18,000 completed questionnaires.

The PROFIEL studies can make their selections because of the close cooperation with the Netherlands Cancer Registry. This registry registers all new diagnoses of cancer in the Netherlands, leading up to over 2 million registered cancer diagnoses.

Important elements of this infrastructure are its open access (to any academic researcher, both in the Netherlands and abroad) and its population-based (with regard to cancer) nature.

Data Quality

The team of PROFIEL is in the process of comparing data quality of online questionnaires with paper questionnaires. Within the existing PROFIEL studies panel response experiments are performed to measure this. Futhermore, because we cooperate intensively with CentERdata, we can rely on their knowledge about data quality. At CentERdata several studies on data quality have been carried out in the LISS panel.

Procedure

The PROFIEL studies panel was set up to accommodate quality of life studies performed by PROFILES itself (research group of Tilburg University and Netherlands Comprehensive Cancer Organisation). After a panel was build and the system had proved to be successful, it became accessible for external researchers.

Academic researchers worldwide can submit a research proposal for data collection in the PROFIEL studies panel. The proposals are evaluated by the PROFILES Project Manager on scientific quality and relevance. More detailed information on the procedure is available at the following sources:
http://www.profilesregistry.nl/Submit

Once a research proposal has been accepted, a PROFILES staff member 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. PROFILES has the right to revise or decline questions which it deems unsuitable for the panel members.

Prior to data collection and the start of the project, the Client Researcher is required to sign an agreement from PROFILES/Netherlands Comprehensive Cancer Organisation.

The PROFILES Project Researcher is responsible for the correct collection and processing of the data. After completing the field work, the PROFILES Project Researcher delivers the data to the Client Researcher.

Recently a tool has been developed which gives the external researcher direct access to the raw data. At the end of the data collection period an external researcher will receive a complete and checked data set.

6.2 Ingestion

Once a questionnaire has been fielded in the PROFIEL studies panel, the concerned PROFILES Project Researcher processes the raw data into an SIP (Submission Information Package). All data processing steps are documented in and run using an SPSS syntax file to ensure a full audit-trail to the original data file and a reconstruction of the data processing.

To prepare the SIP, PROFILES Project Researchers follow a procedure which is documented in the form of a codebook, containing data and metadata requirements. For each SIP, there is an internal second-reader check before the SIP is delivered to the Data Archive Employee.

After the Data Archive Employee receives the SIP, he checks the concerned data and metadata before ingestion in the PROFILES Registry Data Archive. Before the SIP is converted into an AIP (Archival Information Package) and accepted into the PROFILES Registry Data Archive, the Data Archive Coordinator follows a Data Entry Checklist, which defines the required checks on the submitted data and metadata.

The data that are archived in and disseminated via the PROFILES Registry Data Archive are also deposited in the EASY online archiving system of DANS. These data are systematically entered into the EASY system by a designated Data Archive Employee of PROFILES. Once these data have been uploaded to the EASY system, a designated DANS employee verifies the data and if necessary will contact the PROFILES 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 PROFILES Registry Data Archive for the actual data files.

6.3 Archival Storage and System Architecture

PROFILES uses a system which is based on Questasy from CentERdata. This system is the technical basis of the PROFILES Registry 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 to 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. This prompted CentERdata to build Questasy; for more information visit the DDI Alliance website (http://www.ddialliance.org).

Questasy is a web application built using a PHP framework that uses a relational database to store data. The LISS Questasy server is also harvestable using an OAI-PMH implementation. Questasy source code is available for free 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

6.4 Data Management and Administration

Within the context of the OAIS model, data management and administration includes information on database requests and events as well as related statistical information required by the archive administration and management. Also information on customer profiles and preservation process history is included, enabling tracking the migrations of AIPs and including media replacements and AIP transformations.

Administrative information on database events and requests are logged by the application and can be used to verify past events. To access the data archival system (PROFILES Registry), 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, PROFILES staff members need to register to access the system and, 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 action within the system is logged and can be traced back to the individual user.

The data that are deposited in the PROFILES Data Archive are collected within the PROFIEL study panel. When data files are created at the end of the data collection process, all data processing steps are documented in SPSS syntax files and/or Word 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. This procedure, including the file name that is saved, is included in the syntax files.

As part of the SIP, a metadata document referred to as 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 applies. The original SIP is modified by the PROFILES Data Archive Employee, using the same documentation procedure as for the first version, i.e. a syntax file is created for the data file including the modifications of the data file. A new version number is allocated to the file. If 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 changed. The changes in the data are documented in a Word document, which is saved in the same internal working directory as where the SIP is stored. After this the Data Archive Employee delivers a new SIP

version to the Dissemination Manager along with the Word file. The Dissemination Manager enters the new version of the data into the data archive and enters information on the modifications, documented internally in the Word 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 any single moment.

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 the data file by recalculating the checksum of the current files on the server and comparing those values with the checksum determined during upload of the file. These checksums are currently calculated by the system but not displayed externally, by default. 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 PROFILES data is simple and open to every academic researcher, both in the Netherlands and abroad. In principle one year after delivery to the original Researcher, the data are made available by PROFILES to scientific researchers through the PROFILES data website: http://www.profilesregistry.nl.

An extensive set of metadata on the whole life-cycle of the research project are freely 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. Information on publications related to the data is provided as well, if available. The website offers several ways to search 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, 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 Data of the PROFILES Registry, available by sending a request e-mail to the Dissemination Manager.

The signed statement is verified by the Dissemination Manager and approved by the Executive Team. The Dissemination Manager sends the login information by e-mail following access approval. The Data User can then download all published datasets within the database.

To enable users to harvest the metadata of the PROFILES Data Archive, our repository supports the OAI-PMH protocol (base-url is https://cdata8.uvt.nl/oai-server/ikz/oai2.php). Dublin Core metadata information about published study units can be harvested here. The PROFILES Data Archive metadata can also be searched by Google.

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

6.6 Preservation Planning and Long-term Preservation Strategy

The 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 climate controlled professional server rooms.

Preservation ('planning functional entity') is secured further by backing up the data. All servers on which PROFILES data are stored are backed up daily. The backups are encrypted and stored at a different location.

Since the data submitted to the PROFILES Data Archive is stored by CentERdata, the Ingest functional entity is integrated in the systems of the archive. This backup is made by VANCIS, the Dutch super computer center.

A System Administrator of CentERdata 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, PROFILES also 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 via the PROFILES Data Archive only, PROFILES is implementing a Statement of Intent with DANS to grant access via the EASY system, in case the PROFILES Data Archive should ever collapse. While the primary goal is to guarantee long-term preservation through good management of the PROFILES Data Archive, this additional measure serves to create maximum trust in long-term preservation.

Currently, DANS creates persistent identifiers, in this case URNs, for the PROFILES data files when they are ingested by the EASY system. These can be viewed on the website of the EASY system.

7 Data Safeguarding

7.1 Security and Risk Management

All data in the PROFILES Data Archive 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.

7.2 Media Monitoring and Refreshing Strategy

All data in the PROFILES Data Archive 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, 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 it 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-threshold 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 1 February 2016 from http://www.ddialliance.org/what/.

NWO (2006). An Advanced Multi-Disciplinary Facility for Measurement and Experimentation in the Social Sciences (MESS). Retrieved on 1 February 2016 from http://www.nwo.nl/onderzoek-en-resultaten/onderzoeksprojecten/i/45/2445.html

NCDD (2013). Website Netherlands Coalition for Digital Preservation (NCDD). Information retrieved on 1 February 2016 from http://www.ncdd.nl/blog/?page_id=447.

Open Archives Initiative (2013). Website of the Open Archive Initiative. Information retrieved on 1 February 2016 from http://www.openarchives.org/pmh/.

VSNU (2005). Gedragscode voor gebruik van persoonsgegevens in wetenschappelijk onderzoek. Retrieved on 1 February 2013 from http://www.vsnu.nl/code-pers-gegevens.html.