Data Management Plan

Mark Sanders, Martina Chýlková

Document Identifier

D1.9 Data Management Plan

Version

1.0

Date Due

M6

Submission date

30 November, 2015

WorkPackage

WP1 Management and coordination

Lead Beneficiary

UU

Grant Agreement Number 649378

Change log

Version	Date	Amended by	Changes
1.0	23 November 2015	Martina Chýlková	- Document finalized for submission

Content

Executive summary

The purpose of this document is to describe the data management life cycle for all data sets that will be collected, processed or generated by the FIRES project. This document provides a general overview of the nature of the research data that will be collected and generated within the project and outlines how these data will be handled during the project and after its completion. This first version of the DMP serves as starting point and guidelines for the researchers in FIRES project. The more elaborated versions will be uploaded in later stages of the project, whenever it is relevant.

1. Prepare

1.1 Data Collection

Databases generated from the project will be submitted to the EC as part of the deliverables planned in the Project:

D3.2 Pan European Database on Related Variety at NUTS-2 level

D4.2 Pan European Database Time Series GEDI at National Level

D4.4 Pan European Database REDI at Regional Level

D5.1 Database on Start up Processes

Data necessary for these deliverables will be collected mainly from public data sources, proprietary and public sources and through surveys.

In particular, data that will be collected/generated in the FIRES project:

Dataset name	Data type	Description of data	Origin/collection source	File Format	Scale
D3.2 Pan European Database on Related Variety at NUTS-2 level	Numerical data at national and regional NUTS-2.	Consists of a number European regions and countries of a certain number of years.	The data will be collected from different sources of which the GEM and the Skill-relatedness data of Neffke & Henning (2013) are two.	STATA (.dta)	Not known yet.
D4.2 Pan European Database Time Series GEDI at National Level	Numerical data at national level from 2002 to 2014	The database includes institutional and individual indicators that characterize the national system of entrepreneurship and refer on the performance of entrepreneurships in the involved countries.	Individual data: GEM; institutional data: various sources (World Economic Forum, UN, UNESCO, Transparency International, Heritage Foundation/World Bank, OECD, KOF, EMLYON Business	Excel (xlsx)	7,5 MB

			School, IESE Business School). As compared to previous GEDI data collection the Coface risk measurement has been replaced by OECD indicator. Owner: GEDI		
D4.4 Pan European Database REDI at Regional Level	Numerical data in NUTS-1 and/or NUTS-2 (if feasible – requires sufficient sample size)	This only refers to the entrepreneurship indicators that feed into REDI. Approximately 125 region cells for two time periods: 2007-2011 and 2012-2014. In case this is not feasible: 125 regions for one time period: 2010-2014	Researchers are members with GEM and have access to the data	.xlsx (Excel) and .dta (Stata)	Limited size
D5.1 Database on Start up Processes	Mostly quantitative (numerical) data and some qualitative (interview quotes) data at corporate level that can, inter alia, be sorted by country and industry (via NACE, NAICS, US SIC codes)	Venture creation processes of 800 start-up companies in the US, UK, Germany and Italy. Dataset is restricted to alternative energy and ICT companies. The sample is based on external database Orbis.	via CATIs with support of external call center. UU will be the owner	.xlsx / .sav	60 MB

Requirements for access to existing datasets (previously collected data):

Dataset name	Description/summary	Data owner/source	Access issues (requirements to access existing data)
Global Entrepreneurship Monitor (GEM)	Data based on adult population surveys to adult population in European countries	GEM	GEM members (including some FIRES members) have access to the micro data, regional indicators can be compiled and published on mutual consent of the GEM National Teams concerned
Perfect Timing (PT) Database	Venture creation processes of 420 start-up companies in the US, Germany and the Netherlands.	Utrecht University: Andrea	PI (Andrea Herrmann) is the owner of the data

	Dataset is restricted to alternative energy and ICT companies. The sample is based on external database Orbis.	Herrmann	

1.2 Data Documentation

The aim of the FIRES project is to document data in a way that will enable future users to easily understand and reuse it.

All Datasets are Deliverable as a data file and will be labelled with a persistent identifier received upon depositing the dataset. To all datasets, there will be a separate report provided, describing in detail the collection and presenting the descriptive statistics and data manipulations of each data series in the dataset; and will be stored alongside the data.

Common ***metadata*** that apply to all studies in your FIRES project on study level will include. i.e. name, description, authors, date, subproject, persistent identifier, accompanying publications, etc. For such generic metadata the Dublin core or DDI metadata standard will be used. For D3.2 a new metadata template must be developed; D4.2 and D4.4 can follow practice developed in the GEDI- and REDI-indicators; whereas D5.1 can rely on earlier work by Dr. Andrea Herrmann in her earlier Marie-Curie project, where she collected exactly the same type of data in Germany and the US.

File naming and folder structure:
In order to better organize the data and save time the file naming convention will be used to enable titling of folders, documents and records in a consistent and logical way. The data will be available under filename composed of the project Acronym and the Deliverable number, for example: FIRESProjectD32.dta, FIRESProjectD42.dta, FIRESProjectD44.dta and FIRESProjectD51.dta. reports will be stored under corresponding names. Furthermore, specific project/data identifiers will be assigned. All variables are given logical three letter codes and a complete codebook is provided, with definitions and descriptive statistics.

2. Handling research data

2.1 Data Storage and Back-up

Raw data will be stored on secure university fileservers and back up versions will be saved on external portable storage devices (CD) and on personal computers of responsible researchers.

For the duration of the project the research data ***master files*** will be stored on the university fileserver with the partner institution of the responsible PI in order to ensure long term a and secure storage. From the master file location, ***backups*** will be made and stored on local drives – on personal laptops with responsible researchers. ***Working copies*** will be accessible on cloud storage (Dropbox) that enables researchers to access the data and allows editing environment. The updated working copies will be synchronized regularly (after every edit) with the *master copy* location. The person responsible for the synchronization will be the responsible researcher (the researcher who is responsible for generating the data, i.e. Deliverable coordinator).

***Version control**:* Both master copy and back up versions will be using the same identifier for newer versions to ensure the authenticity of the data and to avoid work with outdated versions of files. For different versions codes will be used: V1.00, V1.01; V2.01 etc. with ordinal numbers indicating major and decimals minor changes. The original and definitive copy will be retained. During the research also the intermediate major versions will be retained to make it possible to go back in versions if needed.

STATA also allows for do-files that code all manipulations in the data. All data sets generated in STATA will be thus presented as a collection of *raw source files* (with reference) and a series of *.do-files* that allow for exact replication of aggregation, manipulation and analysis of the data. These .do-files are published with the raw and final cleaned data files.

2.2 Data Access and Security

Within the duration of the project only the directly responsible researchers have *access* to the data files. They are thereby also responsible for the integrity of the datasets and required to carefully document collection and any manipulations made to the data. Data will be made public only after publication of the reports and deliverables. For privacy reasons, raw microdata in D5.1 will remain restricted access after the project, as do the proprietary parts of the data used in D4.2 and D4.4. We will publish data required for the reproduction of analyses. Principal investigators will control the data up to the delivery of the deliverables.

Ownership of the data generated in the project lies with the beneficiary (or beneficiaries) that generates them, as stated in the FIRES Consortium Agreement. In case of joint owners of the data, these shall agree on all protection measures of the data.

The data collected through survey in D5.1 will be anonymized. No privacy /sensitive data are involved in the project.

3. Preserve and Share

3.1 Data Preservation and Archiving

All data generated by the project should be preserved permanently. They will be preserved in Stata .dta and .do as well as a simpler database formats. Together with the data also reports in .pdf and STATA .do-files will be stored as supportive documentation. For the purposes of long term sustainable archiving of the data suitable archiving system will be chosen in the course of the project.

3.2 Data Sharing and Reuse

Possible audiences identified for reuse of the data are mainly students and scholars. In order to ensure that the data and its metadata can be easily found, reused and cited and can also be retrieved even if at some point its location changes, all data generated from the project will be deposited in a public research data repository. Suitable repository that allows the assignment of a persistent identifier as well as for long term storage and open access, will be chosen through re3data.org - registry of discipline-specific repositories. In order to create clarity for potential users towards the use of the data, suitable licenses will be assigned to the data, using creative commons licenses (mostly CC-BY).

Once delivered to the European Commission and approved, the data files will also be made public on the website of the project. The data for deliverables D4.2 and D4.4 are proprietary, but aggregated data can be made public. Micro-data for D5.1 will not be made public until all reports foreseen in the project have been published.