Grant agreement No. 822390

MICROPROD

Raising EU Productivity: Lessons from Improved Micro Data

H2020-SC6-TRANSFORMATIONS-2018
Supply and demand-oriented economic policies to boost robust growth in Europe – Addressing the social and economic challenges in Europe

D1.1
Data Management Plan

WP 1 – Firm-level Data and Productivity Measurement

Due date of deliverable	30/06/2019 (Month 6)
Actual submission date	28/06/2019 (Month 6)
Start date of project	01/01/2019
Duration	36 months
Lead beneficiary	IWH
Last editor	Mirja Hälbig
Contributors	Eric Bartelsman, Mirja Hälbig, Filippo di Mauro, Evghenia Scripnic

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 822390.

Disclaimer

The content of this deliverable does not reflect the official opinion of the European Union. Responsibility for the information and views expressed herein lies entirely with the author(s).

History of the changes

Version	Date	Released by	Comments
0.1	24-06-2019	Mirja Hälbig	First draft
0.2	25-06-2019	Evghenia Scripnic	Input for ethics section
1.0	27-06-2019	Filippo di Mauro	Final version

Table of contents

Key word list

Data, Sources, Tools, Format, Standards, Quality control, Reproducibility, Metadata, Identifiers, Storage, Security, Share

Definitions and acronyms

Acronyms	Definitions
BR	Business Register
SBS	Structural Business Survey
IS	Community Innovation Survey
EC	Community survey on ICT usage and e-commerce in enterprises
LP	Long Production Panel
EX	Export
IM	Import
EE	Linked Employee-Employer Data
FD	Financial Data
RTD	Research and Technical Development
EDI	Electronic Data Interchange
CRM	Customer Relationship Management
ADE	Automated Date Exchange

1. Introduction

Data management is a key task for all data producers and/or data re-users working within the framework of the Horizon 2020 research program. Within the MICROPROD project, Data Management Plans (DMPs) is due by Month 6 (30th June 2019), and it will be updated whenever significant changes arise (e.g. in consortium policies, consortium composition or external factors) or when new research data should be collected.

One goal of the DMP is to identify specific types of data to be handled within MICROPROD in close collaboration with the consortium partners. As a result, the identification of specific types of data also allows determining where Informed Consent Forms (ICF) and Data Protection Certificates (DPC) are needed and who will be responsible to collate those documents.

The DMP includes information and procedures specifying how research data will be organised within MICROPROD (during and after the project) and how to ensure its curation, preservation and sustainability. It is not only intended to show how data can be effectively used and re-used but also how data should be published and archived (what parts of research data will be open and how). Data management needs specific knowledge about the type of research data and the data lifecycle as well as the data collection mechanisms. The DMP does not contain the specific procedures for achieving research results. The DMP evolves and gains more precision and substance during the lifespan of the MICROPROD project.

The DMP only considers **research data** that are specifically created during the research process for the purpose of analysis and to produce research results, by using a scientific methodology and by MICROPROD beneficiaries. This data might include observation data, survey data (e.g. by interviews), statistical data. Text publications as one result of the research process are not considered as research data to be treated within the DMP.

In general, different steps for getting data “fit for use” are necessary as data quality is a main challenge in every data collection and processing task. This particularly includes the elements for checking consistency and accuracy, error detection, data cleaning (including a script of the cleaning process), versioning and documentation (e.g. by quality flags, metadata, coding)[1].

The DMP will be used by consortium partners to describe how data is being collected / accessed / managed / made available by the project.

The DMP will:

- Detail the specific data to be collected, handled and processed by the project;
- Clarify the responsibilities among project partners about the management (collection, storage, access, processing) of the identified data;
- Refer to ethical aspects in terms of processes and the rules to be adopted by each partner to comply with the requirements (national law and the EC Data Protection Directive, etc.) identified in the Ethics Deliverables;
- Identify data which can be openly published / made available by the project;

[1] German Federation for Biological Data (2019). GFBio Training Materials: Data Life Cycle Fact-Sheet: Data Life Cycle: Assure. Retrieved 25 June 2019 from https://www.gfbio.org/training/materials/data-lifecycle/assure.

As a particular interest, data management within MICROPROD is intended to describe accessibilities to data for beneficiaries and other potential users (i.e. researchers, public). Since for H2020 projects specific repositories for data storage are not imposed, MICROPROD data will be eventually deposited in the locations, which will be individually judged as the most appropriate in due course.

The DMP will address the relevant aspects of making data FAIR (findable, accessible, interoperable and re-usable), including what research data the project will generate, whether and how it will be made accessible for verification and re-use, and how it will be curated and preserved.

Information on the procedures for data collection, storage, protection, retention, and destruction, and confirmation that they comply with national and EU legislation are also a component to be included in the DMP.

To ensure the transparency of the research data it is strongly recommended to document all important data edits and to report those activities at the DMP.

2. Data Summary

This section contains the summary of research data collected in MICROPROD with particular attention to the purpose of data collection/generation and the relation to the MICROPROD objectives (including the outline of data utility). Specifying the concrete types and formats of data, their origin and expected size is not feasible at this very early stage of the project. These components will be specified during the project by updating the DMP as a living document.

In general, research data can be categorised into three basic types[2]: Raw Data (initial, unprocessed data), Primary Data (processed (raw) data, and Secondary Data (re-used (primary) data originally collected for other purposes). Data will be collected across many activities, tasks and subtasks throughout the whole lifespan of the MICROPROD project.

There are several types of research data in MICROPROD that need to be carefully managed. The data will be collected and produced mainly within WP1 with the purpose to – among others - equip the other work packages with an extended database. This particularly includes the following topics:

Funded within the framework of the EU Research and Innovation program “Horizon 2020”, MICROPROD aims to (i) expand and deepen the data and methods available for measuring productivity, (ii) generate new insights into the causal mechanisms of productivity developments, and (iii) make them publicly available.

MICROPROD intends to expand the micro-data infrastructure available for productivity measurement, in the context of its Work Package 1 (WP1). In close cooperation with selected National Statistical Institutes (NSIs) serving as “Pilot” Institutions, we will collect and harmonize information (meta data) from confidential firm-level data from these national

[2] See: German Federation for Biological Data (2019). GFBio Training Materials: Data Life Cycle Fact-Sheet: Data Life Cycle: Analyze. Retrieved 25 June 2019 from https://www.gfbio.org/training/materials/data-lifecycle/analyze.

sources. The meta data to be collected stems from registers maintained by national statistical institutes, such as the Business Register, and from EU harmonized surveys, implemented by national statistical institutes, such as the Structural Business Survey (SBS), the Community Innovation Survey (CIS), the E-Commerce Survey (EC), but also includes registers or national surveys regarding trading activity, linked employee-employer data and firm financial data.

Our research will contribute to better productivity measurements. By the end of the project, we will have created the means for harmonized cross-country data analysis in WP1 that is relevant for measuring the factors that drive productivity. In particular, we aim at overcoming the current problems affecting productivity measurement at the aggregate level by exploiting the availability of detailed firm level data, linked with information on employees, intangibles, and other indicators of firms' performance.

The insights obtained could be used to propose a set of micro-based moments that can be collected in an open source environment at a level of aggregation (to maintain confidentiality) to be determined. This work could for example, lead into a proposal for indicators to be included in the Competitiveness Research Network (CompNet) dataset.

The data infrastructure will be composed of four components:

- A metadata catalogue of firm level information available in selected Pilot NSIs;
- A set of aggregated information (moments and joint distributions) at the sector level of variables related to productivity indicators and possible drivers (finance, labor, trade, competition and alike)
- The novel part of the project, however, will be to allow researchers to run their own code on a set of variables harmonized across countries (in terms of definition, sample characteristics etc.) in close cooperation with the NSIs.
- The findings of the pilot countries can be used as blueprint for other national statistical offices to participate in building a data infrastructure.

The data infrastructure will allow future researchers to have easy access to a larger set of productivity related data in a selection of countries.

3. FAIR Data

3.1. Making data findable, including provisions for metadata

Standard identification mechanisms are strongly recommended wherever possible. Persistent and unique identifiers are typically used in research data repositories (a persistent identifier (PID) is for instance the DOI – Digital Object Identifier). Whether collected research data can obtain persistent identifiers depends from the type of collected data and privacy issues. It is also strongly recommended to define or at least suggest how specific research data should be cited. Persistent identifiers help researchers to make their research data traceable and better citable.

Naming conventions help to systematically store files created within the MICROPROD project. The Data Manager hereby outlines some basic conventions for making data easily findable.

Depending on the particular purpose (e.g. data exchange among beneficiaries or data archiving/storage) it is most suitable to name files starting with the date in ISO data format "year-month-day" with a 4 digit years and hyphen between the data components. It is also most recommendable avoiding blank characters and special characters as well as very long file names. Instead, an underscore character should be used to separate denominations within a file name. It is also helpful indicating the type of content (e.g. Deliverable (D), Presentation (PRES), and Calculation (CALC)). Finally, a short description of the content (after the last hyphen) is also recommendable. Version numbers are also useful as far as different versions of the same document exist.

Search keywords will be provided for all MICROPROD research data sources by each responsible partner that optimises possibilities for re-use. Specific keywords and keyword conventions will be established as soon as research data are collected. The Data Manager encourages search engine optimization measures as part of a dissemination strategy.

Managing change and version control (including clear versioning) is foreseen in all internal and public documents. Deliverables will get a version number on the cover page. This number is being continually updated during the document creation process. Other data and documents should include a "Timetable For Updates" containing the history of changes with version, publication date, autor, change description.

Provision of metadata is an important task during the whole data collection process. Beneficiaries should be aware of that completing this task is paramount for giving the opportunity to use the produced research data by others. Metadata should carefully be produced and stored (either embedded or in a separate data file) during and after the project period depending on the specific data content. The content of metadata strongly depends on the data format (type of research data). For some research data (i.e. survey data), an additional codebook containing all variables, descriptions, values and meanings is strongly recommended.

We strive to ensure that the collected metadata complies with the SDMX standards.

Typical elements of metadata are information about data classes, data properties and the encoding schemes. The minimum of metadata information includes the following items: creator of the data, type of study or data (content of the data), methodology for data generation, data format, detailed description of variables (if any) or records, date of origin/recording or year of publication, location of data collection.

This metadata information helps beneficiaries and further data users to discover research data of the MICROPROD project. Project partners are encouraged to describe these contents clearly and understandable so that people without specific knowledge can also catch the metadata information. Opportunities to convert metadata into a compatible standard afterwards will be checked depending on the amount and type of research data finally compiled within MICROPROD.

Work Package 1 data infrastructure will be findable through a metadata catalogue, which will contain information on data availability by country. This will include information on the sample frame and unit of observation of the data sources as well as harmonized definition of variables in each data source.

Meta data will be identified using standard identifiers and adequate links will make such information available in machine-readable form for users.

3.2. Making data openly accessible

MICROPROD will widely share the results to improve the statistical infrastructure available for researchers and policy makers. This infrastructure will build on the ideas and methods explored in previous ESSnet projects and as implemented in CompNet.

Further, the micro-aggregated moments used in the MICROPROD work packages will be made available at a level of aggregation to be decided in light to preserve data security. The source datasets itself will not be available. However, researchers will be able to run their own codes on a previously harmonized set of variables via remote access or execution at the NSIs in accordance with their respective rules.

Software to be used is likely to be STATA and/or R

Metadata will be deposited in a common accessible server. Underlying micro-level source data will not be exchanged.

Access to the underlying micro-level data through the research infrastructure will be regulated with remote access agreements with the individual NSIs.

3.3. Making data interoperable

The MICROPROD project consortium will strive to producing data in an interoperable form that allows data exchange and re-use between researchers, institutions, organisations, and countries. All processed information will be provided in English.

Interoperability will also be insured via harmonizing in full detail and across countries data definitions and indicator construction.

A data user manual will provide information on variable definition and other relevant methodologies related to indicator construction.

3.4. Increase data re-use (through clarifying licences)

One of the objectives of Work Package 1 is to create a document for the EU providing information on how to set up a convenient tool that allows researchers to have an overview of micro-data availability at national statistical institutes, dataset linkability and harmonized variable definitions.

In order to ensure longevity of the project, the insights will be implemented in the CompNet database and made available in the course of the distribution of the respective annual CompNet data vintage.

4. Allocation of Resources

NSIs are currently estimating (as of end June 2019) the actual cost of generating the required metadata. This will be covered using explicit provisions already included in the successful bid.

Other costs to be incurred by NSIs to make the metadata fully FAIR during the project will be included in the next few weeks.

Filippo di Mauro has been appointed Data Manager for MICROPROD.

The additional costs to be incurred by the legacy organisation of the metadata set will be estimated towards the end of the project; coverage of such costs will need to be identified.

5. Data Security

The micro-level data to be used in the research projects is accessible via remote access or execution only at the NSIs in accordance with their respective rules.

6. Ethical Aspects

Data Protection is of special importance in multi-institutional projects such as MICROPROD. Ethical aspects and legal issues of data collection are specifically addressed in Deliverables D9.1, D9.2, D9.3, D9.4, D9.5 and D9.6.

Within MICROPROD's research activities it is envisaged that only previously collected anonymised personal data will be used. This data will be provided by the National Statistical Institutes, which have their own data protection rules, all in adherence to EU and national laws. Personal data held by the NSIs are anonymized. Individual data may not be identified by researcher. All MICROPROD researchers with access to such data at NSIs will have followed all legal and ethical trainings for proper use of such data.

Within MICROPROD's communication and dissemination activities the data minimisation principle will be applied and only standard personal data may be collected (such as names, email addresses, affiliation) but will not be made public unless explicitly permitted by the data subject. All MICROPROD partners involved in collection and processing of personal data will implement informed consent procedures by creating firstly, templates of the informed consent, secondly, by appointing a Data Protection Officer. They will also store the personal data on

secure servers and will not keep it for longer than is necessary for the purposes for which the personal data are processed.

Where any partner deviates from the above described requirements, the Data Manager will bring the matter to the immediate notice of the project coordinator. All partners involved in data collection, storage, and use will be in regular and ad hoc contact with the Data Manager for advice, to ensure that the risks of noncompliance are minimal. This is of special concern since the General Data Protection Regulation (GDPR) has taken full effect in 2018.

7. Conclusions and future steps

Following the first WP1 meeting with Pilot NSIs, the process has been initiated to establish the set of additional variables to be downloaded, homogenized and linked across (see the appendix for a preliminary list). Such process is still ongoing. During the Second meeting, already scheduled to be held in September 2019, such first stage is expected to be completed. At that point, new data will be made available to the relevant researchers in MICROPROD to allow the production of the promised output. An additional task relevant for the DMP will be then to create procedures according to which access to the homogenized data will granted remotely to authorized researchers from the participating NSIs.

APPENDIX: Preliminary List of Variables

File	Variable Description	Type
BR	Unique Firm Identifier	char
BR	year to which data pertain	char
BR	BR closing date	date
BR	NACE1 Orig 4	char
BR	NACE2.2 Orig 4	char
BR	NACE fixed (prepared by NSI)	char
BR	Employment	integer
BR	Foreign ownerwhip (y=1, n=0)	bool
BR	Part of domestic group	bool
BR	Part of international group, domestically owned	bool
BR	Date of commencement of activities	integer
BR	Date of final cessation of activities	integer
BR	NUTS-2 Region	char
BR	Legal Form	code
SBS	Unique Firm Identifier	char
SBS	Year (needed, if data in linked panel)	integer
SBS	Value added at factor cost	euro
SBS	Production Value	euro
SBS	Number of employees in full time equivalent units	integer
SBS	Number of employees	integer
SBS	Number of female employees	integer
SBS	Personnel Costs	euro
SBS	Total purchases of goods and services	euro
SBS	Purchases of energy products (in value)	euro
SBS	Capital Stock (book value)	integer
SBS	Depreciation Cost	integer
SBS	Capital Stock (other measure)	integer
SBS	Dummy for export status (y=1, n=0)	bool
SBS	Sales to foreign (thousand euro)	euro
SBS	Investment in ICT hardware	euro
SBS	Gross investment in concessions, patents, licences, trade marks and similar rights	euro
SBS	Investment in purchased software	euro
IS	Unique Firm Identifier	char
IS	Year (needed, if data in linked panel)	integer
IS	Enterprise part of a group	bool
IS	Country of head office	char
IS	Other EU/EFTA/CC market	bool

IS	All other countries	bool
IS	Introduced onto the market a new or significantly improved good	bool
IS	Introduced onto the market a new or significantly improved service	bool
IS	Who developed these product innovations?	code
IS	Did the enterprise introduce a product new to the market	bool
IS	% of turnover in new or improved products (introduced during xxx) that were new to the market	pct
IS	Introduced onto the market a new or significantly improved method of production	bool
IS	Introduced onto the market a new or significantly improved logistic, delivery or distribution system	bool
IS	Introduced onto the market a new or significantly improved supporting activities	bool
IS	Who developed these process innovations?	code
IS	Engagement in intramural R&D	bool
IS	Type of engagement in R&D	code
IS	Expenditure in intramural R&D (in national currency)	euro
IS	Purchase of extramural R&D (in national currency)	euro
IS	Expenditure in acquisition of machinery, equipment, software (in national currency)	euro
IS	Acquisition of existing knowledge from other enterprises or organisations	euro
IS	All other innovation activities including design, training, marketing, and other relevant activities	euro
IS	Total of the above innovation expenditure (in national currency)	euro
IS	Public funding from local or regional authorities	bool
IS	Public funding from central government	bool
IS	Public funding from the EU	bool
IS	Funding from EU's 6th or 7th Framework Programme for RTD	bool
IS	Cooperation arrangements on innovation activities	bool
IS	New business practices for organising work or procedures	bool
IS	New methods of workplace organisation	bool
IS	New methods of organising external relations	bool
IS	Significant changes to the aesthetic design or packaging	bool
IS	New media or techniques for product promotion	bool
IS	New methods for product placement or sales channels	bool
IS	New methods of pricing goods or services	bool
IS	New or significantly changed sales or distribution methods	bool
IS	During the three years xxx to xxx, did your enterprise apply for a patent	bool
IS	During the three years xxx to xxx, did your enterprise Register a trademark	bool
IS	During the three years xxx to xxx, did your enterprise license out or sell a patent, industrial design right, copyright or trademark to another enterprise, university or research institute	bool
IS	During the three years xxx to xxx, did your enterprise license in or buy a patent, industrial design right, copyright or trademark owned by another enterprise, university or research institute	bool
IS	percent of employees with tertiary degree	pct
EC	Unique Firm Identifier	char
EC	Year (needed, if data in linked panel)	integer

EC	Firm has broadband	bool
EC	Firm orders through computer networks (websites or EDI)	bool
EC	% of orders through internet	pct
EC	Firm sells through internet (or EDI)	bool
EC	% of sales through internet (or EDI)	pct
EC	Firm has internet	bool
EC	% of workers with access to internet	pct
EC	Firm uses computers	bool
EC	% of workers using computers	pct
EC	Firm has website	bool
EC	Firm has mobile access to internet	bool
EC	Enterprise Resource Planning	bool
EC	Use of ADE for receiving e-invoices	bool
EC	Use of ADE for sending e-invoices	bool
EC	Use of ADE for sending payment instructions to financial institutions OR sending transport documents OR receiving data to/from public authorities	bool
EC	Use of ADE for sending or receiving data to/from public authorities	bool
EC	Sharing Supply Chain Management (SCM) data with suppliers	bool
EC	Sharing SCM data with customers	bool
EC	Use of CRM software to share of information with other business functions	bool
EC	Use of CRM software to analyse information for marketing purposes	bool
EC	Firm shared information on sales orders electronically and automatically for management of inventory levels	bool
EC	Firm shared information on sales orders electronically and automatically for accounting	bool
EC	Firm shared information on sales orders electronically and automatically for production or services management	bool
EC	Firm shared information on sales orders electronically and automatically for distribution management	bool
EC	Firm shared information on purchase orders electronically and automatically for accounting	bool
EC	Firm shared information on purchase orders electronically and automatically for management of inventory levels	bool
EC	firm employs IT specialists	bool
EC	firm provides IT training	bool
EC	firm outsource IT functions	bool
EC	firm analyses big data	bool
EC	Big data analysis for the enterprise is done by the enterprise's own employees and by an external provider	bool
EC	Big data analysis for the enterprise is done by an external service provider	bool
EC	Use industrial or service robots	bool
EC	Use industrial robots	bool
EC	Use service robots	bool
LP	Year to which data pertain	char
LP	Unique Firm Identifier	char
LP	Naics code (4 digit)	char
LP	Value added in thousand Euro	euro
LP	Production in thousand Euro (Sales+final inventory change)	euro

LP	Employment (in FTE or persons)	integer
LP	Total labor costs (thousand Euro)	euro
LP	Total intermediate purchases (thousand Euro)	euro
LP	Capital Stock (book value, or describe)	integer
LP	Value added deflator (index 2000=1)	2000=1
LP	Output deflator (index 2000=1)	2000=1
LP	Materials deflator (index 2000=1)	2000=1
EX	Unique Firm Identifier	char
EX	Year to which data pertain	integer
EX	Harmonized product code	char
EX	Value	euro
EX	Destination Country	char ISO
EX	Units	integer
IM	Unique Firm Identifier	char
IM	Year to which data pertain	integer
IM	Harmonized product code	char
IM	Value	euro
IM	Country of Origin	char ISO
IM	Units	integer
EE	Unique Firm Identifier	char
EE	Year to which data pertain	integer
EE	Person identifier	char
EE	Person Age	integer
EE	Gender (m, f)	bool
EE	Education (category from ISO codes)	char ISO
EE	Skill (international classification)	char ISO
EE	Tenure Length (in years)	integer
FD	Unique Firm Identifier	char
FD	Year	integer
FD	Total Fixed Assets	euro
FD	Intangible Fixed Assets	euro
FD	Tangible Fixed Assets	euro
FD	Other Fixed Assets	euro
FD	Current Assets	euro
FD	Cash and Cash Equivalent	euro
FD	Total inventories (ORBIS: Stocks)	euro
FD	Accounts receivable (ORBIS: Debtors)	euro
FD	Other current assets	euro
FD	Total assets	euro
FD	Total shareholder funds (equity)	euro
FD	Non-current liabilities	euro
FD	Long-term debt	euro
FD	Other non-current liabilities	euro

FD	Current liabilities	euro
FD	Short-term debt (ORBIS: Loans)	euro
FD	Accounts payable (ORBIS: Creditors)	euro
FD	Other current liabilities	euro
FD	Turnover	euro
FD	Labour cost (Costs of Employees)	euro
FD	Intermediate inputs (Material Costs)	euro
FD	R&D expenditure	euro
FD	Operating profit/loss (EBIT)	euro
FD	Interest paid and financial charges	euro
FD	Depreciation	euro
FD	Profits and losses before taxes	euro
FD	Cash flow (from profit/loss statement)	euro
FD	Dividends	euro