The Granular Economy

A Data Management Plan created using DMPonline.be

Creators: Anneleen Fastenaekels, Jo Reynaerts

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Principal Investigator: n.n. n.n.

Data Manager: Jo Reynaerts

Project Administrator: Anneleen Fastenaekels

Grant number / **URL:** METH/21/01

ID: 197572

Start date: 01-10-2022

End date: 30-09-2029

Project abstract:

This project aims to provide a new and different approach to analyze macroeconomic performance, by recognizing that heterogeneous behavior of the underlying microeconomic parts provide important new insights about the channels of macroeconomic fluctuations. For instance, a country which has all the same size of companies, is likely to respond differently to an external economic shock (such as COVID-19) compared to a country with many small firms and a few large firms. The reason is that very large companies often play a central role in the production network, consisting of a lot of small firms which are linked to these large companies through the supply chain. Thus a shock affecting a large central firm ripples through the entire production network affecting small firms indirectly. This approach is sometimes also referred to as the 'granular origins' of aggregate fluctuations. The program is structured along various micro data sets that have been developed over the past years. As such, we aim to identify the 'DNA' of an economy. To this end, the program will engage in unravelling the micro channels affecting the key macroeconomic indicators commonly used in economic policy: inflation, GDP growth and unemployment. More specifically the program aims to (i) prices and inflation: understand the micro channels affecting the pass-through from international shocks to domestic prices and understand the role of price setting and price heterogeneity for aggregate price fluctuations and inflation; (ii) productivity and GDP growth: determine the importance of demand-side factors in explaining the heterogeneity and persistence in revenue productivity and understand its relationship with aggregate productivity growth; and (iii) firm level employment growth and unemployment: understand how firm heterogeneity matters for aggregate employment and unemployment fluctuations, in particular in the context of recessions, trade wars (US, China), trade shocks (Brexit, Trumpet), and negative supply-side shocks (COVID-19).

Last modified: 31-03-2023

The Granular Economy

Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume	Dimensions	Coverage	Source	Type	Access
		*Indicate: **N**(ew data) or **E**(xisting data)*	Indicate: **D**(igital) or **P**(hysical)	Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA						
WIOD	World Input-Output Database	E	D	N	.csv, .RData	< 1GB	NA	43 countries, 56 sectors	balanced panel, yearly, 2000-2014	Rijksuniversiteit Groningen	open source	https://www.rug.nl/ggdc/valuechain/wiod/
Prodcom	Production of manufactured goods	E	D	N	.csv	< 1GB	NA	8-digit product level, country level (EU)	unbalanced panel, yearly, 1995-2021	Eurostat	open source	https://ec.europa.eu/eurostat/web/prodcom
Comext	Internationational trade data	E	D	N	.csv	< 10 GB	NA	imports and exports at 8-digit product level (CN8), country-pair level	unbalanced panel, monthly, 2001-2021	Eurostat	open source	https://ec.europa.eu/eurostat/web/international-trade-in-goods/data/focus-on-comext
Comtrade	Internationational trade data	E	D	N	.csv	< 1TB	NA	imports and exports at product level (6-digit HS, 5-digit SITC, 3-digit BEC), country-pair level	unbalanced panel, monthly, 2000-2021	United Nations	open source	https://comtradeplus.un.org/
QED	Qualitative employment data	E	D	N	.csv	< 1GB	NA	employment data in terms of workers, number of hours worked, compensation of employees and self-employed workers, by industry (NACE A38), accupational status, gender, age class, and educational level	balanced panel, yearly, 1999-2020	Federaal Planbureau	open source	https://www.plan.be/databases/data-23-nl-kwalitatieve_werkgelegenheidsdata_voor_belgie_1999_2020
LFS	Labor Force Survey	E	D	N	.csv	1 GB	NA	individual worker level type of contract, sector of employment, labor income by income decile for employees, additional covariates	cross-section, quarterly, 1999-, by income decile	Statbel	open source with application procedure	https://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey
BEL-FIRST	Firm-level accounting data	E	D	N	.csv	< 10 GB	NA	annual accounts data for over 1.2 million Belgian firms including a.o. employment, added value, turnover, wages	unbalanced panel, yearly, 2000-2021	Bureau van Dijk	proprietary	https://www.bvdinfo.com/en-gb/our-products/data/national/bel-first
MRIO	Multi-Regional Input-Output Tables	E	D	N	.csv	< 1GB	NA	143 sectors over 3 Belgian regions	cross-section, 2015 (ESR 2010)	Federaal Planbureau	proprietary	Prof. dr. Joep Konings
NBB 1	Balance sheet information	E	D	N	OBDC, .dta	< 10 GB	NA	firm-level financial and operational indicators	unbalanced panel, yearly, 1984-2021	Nationale Bank van België	proprietary	Prof. dr. Joep Konings, Prof. dr. Hylke Vandenbussche
NBB 2	Firm-level trade data	E	D	N	OBDC, .dta	< 10 GB	NA	firm-level imports and exports at 8-digit product category, by origin-destination pair	unbalanced panel, monthly, 2005-2021	Nationale Bank van België	proprietary	Prof. dr. Joep Konings, Prof. dr. Hylke Vandenbussche

BTW 1	VAT declarations	E	D	N	OBDC, .dta	< 10 GB	NA	VAT declarations by VAT-liable firms (turnover, expenditures, investment)	unbalanced panel, monthly/quarterly, 2005-2021	Nationale Bank van België, FOD Financiën	proprietary	Prof. dr. Joep Konings, Prof. dr. Hylke Vandenbussche
BTW 2	VAT listings	E	D	N	OBDC, .dta	< 10 GB	NA	point-to-point listings of sales value by VAT-liable firms	unbalanced panel, yearly, 2002-2021	Nationale Bank van België, FOD Financiën	proprietary	Prof. dr. Joep Konings
RSZ	Employment and wages	E	D	N	.csv	< 10 GB	NA	establishment-level employment and wages	unbalanced panel, quarterly, 2016-2021	Rijksdienst voor Sociale Zekerheid	proprietary	Prof. dr. Joep Konings
VKBO	Establishment-level information	E	D	N	TBD	< 10 GB	NA	establishment-level information on location, NACE codes, and status of activity (normal, failure, M&A, ...)	one wave (2016)	FOD Economie, EWI	proprietary	Prof. dr. Joep Konings
Scanner EU	Household scanner (expenditure) data	E	D	N	.csv	~ 100 GB	NA	8 countries, 3,500-40,000 households, 1B transactions	daily, 2010-2020	GfK/Kantar	proprietary	Prof. dr. Frank Verboven
Scanner US	Household scanner (expenditure) data	E	D	N	TBC	TBC	NA	50,000 households	daily, 2010-2020	AC Nielsen	proprietary	Prof. dr. Jan De Loecker
Scanner KZT	Store-level scanner (expenditure) data	E	D	N	.csv	< 1 GB	NA	12 store types, top 40 SKUs across 80 categories	monthly, 2014-2016 and 2019-2021	AC Nielsen	proprietary	Prof. dr. Joep Konings
Scanner Metro	Store-buyer-level scanner (expenditure) data	E	D	N	.txt	< 10 GB	NA	NA	daily, 2014-2017	Metro	proprietary	Prof. dr. Anatoli Colicev
Automobile EU	Product-level sales, price and characteristics data	E	D	N	.csv	2 GB	NA	7 countries, 20 years, > 10k products	annual, 1998-2018	JATO	proprietary	Prof. dr. Frank Verboven
COVID-19 support measures	Firm-level information on COVID-support	E	D	N	.csv	< 1GB	NA	firm-level, one region, by year	unbalanced panel, 2020-2022	Vlaams Agentschap Innoveren en Ondernemen	proprietary	Prof. dr. Joep Konings
Relance programs	project-level information on federal and regional relance measures	E	D	N	.csv	< 1GB	NA	project-level, by region and cluster, by year	pooled cross-section, 2020-2022	Federaal Planbureau	proprietary	Prof. dr. Joep Konings
Distances	ZIP-to-ZIP travel distance and travel time data	E	D	N	.csv	< 1GB	NA	NA	500,000 ZIP-to-ZIP observations	Localyze	proprietary	Joris Hoste

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

We refer to the table above for data descriptions, identifiers, sources, and access details.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- No

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- Yes (Provide PRET G-number or EC S-number below)

Most of the firm-level data sets and all the household surveys are to be considered as personal data as they contain confidential information. The Labour Force Survey is pseudonymized; the pseudonymization is done by the data provider (Statbel).
Confidential firm-level information is included in the (VAT-based) transactions data from the Nationale Bank van België (NBB) and FOD Economie, firm-level financial accounts from Bureau van Dijk and NBB, and the establishment-level information on employment and wages from RSZ. Information on self-employed entities can be protected by GDPR rules. Depending on the dataset, these observations are not accessible to us. For each of the firm-level data sets as well as the LFS, access and permission to handle will be assigned to specific researchers on a case-by-case basis by the work package leaders (see also below). The use of the data is bound by contractual obligations stipulated by the supplier of the data, and the internal rules of the KU Leuven (www.kuleuven.be/privacy). The data will be recorded in the respective institutions' privacy registers (see more details below).

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or Data transfer agreements, Research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

Both the firm-level annual accounts, trade and VAT-based transaction data (NBB and Bureau van Dijk), as well as the establishment-level wage and employment data (RSZ) used for this project are subject to restrictions, only allowing us to disseminate statistics at the more aggregated sector-region level.
Dissemination or use of the survey data provided by Statbel and ECB/NBB is dependent on the contractual agreement between the researchers and Statbel or ECB/NBB. For LFS, the contractual obligations are general: individual records cannot be revealed. Sharing of the dataset or intermediate results at the micro-level between different research institutions is not allowed. However, aggregates and joint moments of the data can be shared. On the production side, statistics at the level of detailed NACE codes (143 sectors) and over time, will be provided if and where applicable.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, codebook.tsv etc. where this information is recorded).

In agreement with current best practices put forward by leading economic journals, metadata for each dataset used in the various work packages of the project will include a full description of the data set at hand, its coverage, variable names and descriptions, format and disk size. Additionally, working papers resulting from the work packages will contain a technical appendix detailing all data sources, data handling, variable creation. Upon completion of the project, all data (with the exception of proprietary data), computer code, logbooks, resulting graphs and tables will be made publicly available. All metadata will be accompanied by a corresponding README file that includes the former sources of information as well as a description of how to proceed with a replication analysis. More specifically, this file will explain the project folder structure and the set-up of the empirical (when and where applicable). Embedded computer code (State, R, MATLAB, ...) will contain explanatory in-file documentation to explain basic data manipulations as well as estimation procedures and auxiliary output (tables, graphs and figures) generation for other researchers.
More specifically, the following actions will be taken to make the project data FAIR (findable, accessible, interoperable and reusable):

1. Making data **findable**. For each of the datasets used, we will include a full description of the metadata in the coding and data pipelines, and on the project website (probably Github). This metadata includes data source, link to the original datasets (if available), access requirements (if any), coverage, variable names, formats, and total disk size. A description of each variable (units, string/numeric etc.) will be given. We will also provide manuals and related documents wherever available from the providers. As mentioned below, these metadata and the accompanying descriptions will be documented either in the appendix of corresponding working papers, or in specified Technical Notes (TN) stemming from data-related tasks.
2. Making data openly **accessible**. Several datasets are publicly available, and will be made accessible to the general public. Other datasets cannot be made publicly available, as it would be a breach of contract with the respective data providers. In any case, we will provide access to all codes, logbooks and results (graphs, tables, etc.). For all datasets and results, including confidential and pseudonymized datasets, we will provide detailed descriptive statistics on the distributions and correlations of the variables in the dataset, which can be used to generate synthetic data for reproducible research.
3. Making data **interoperable**. Our research teams have extensive experience with confidential datasets and large and complex project pipelines. To that end, we have developed standardized methods to provide accessible and reproducible research outputs. These methods generally follow the recent standards for reproducible research at leading journals such as the *American Economic Review* and the *Economic Journal*, and have been further adapted to ensure the highest possible reproducibility standards. We follow a task-based approach to coding and project pipelines. Data, code and analysis are organized into tasks that break up complex projects into smaller, modular/containerized pieces and that jointly constitute a pipeline into one direction. A task is an indivisible piece, a quantum of workflow. Tasks are sequential, so that the output of one task can be used as the input of another, downstream, task. Tasks are also modular, in the sense that they serve as containers, e.g. to test particular algorithms, sample selections,... and as a natural unit to divide workload among co-authors. Finally, we have dedicated sandbox environments to test and debug code on random data or small subsets of the real data. Results of particular tasks can then be run on the full, real datasets. These sandboxes also serve as a gateway for any interested researcher to check and experiment with the codes and public data
4. Increase data **re-use**. The complete data and coding pipeline, including metadata and support documentation will be published on public repositories on Github, Dataverse, and/or KU Leuven RDR. We will provide all project code in both Stata format and in Python, R or Julia (open access), with explanatory in-file documentation in the Stata do-file or Python/Julia notebooks that explain the data manipulations to other researchers. Related documentation on the datasets will be provided with each dataset, and a detailed description of the nature, manipulations and outputs of each coding task will be documented and shared. This greatly facilitates code sharing between the various project partners/institutions, and for future use by other researchers. For an example of a full pipeline, see https://github.com/glennmagerman/BDMMM_JPE. Conditional on access to the underlying datasets, these repositories allow for a full replication of the projects. Moreover, we will store the full data and coding pipelines also on the respective secure and dedicated servers at KU Leuven. This will allow full replication of all results for interested researchers that have cleared access to the respective datasets. Accessible data will be provided in open formats (e.g., .csv) for use by others.

Will a metadata standard be used to make it easier to find and reuse the data ?
If so, please specify which metadata standard will be used.

If not, please specify which metadata will be created to make the data easier to find and reuse.

- No

For datasets obtained through a non-standardized procedure and for which no full documentation is available on the website of the data provider, a full description will be stored in a metafile, containing their coverage, variable names, units, formats, and total disk size. These will also be documented in detail either in the appendix of the corresponding working paper, or in an accompanying Technical Note (TN).

Data Storage & Back-up during the Research Project

Where will the data be stored?

- Shared network drive (J-drive)

In general, data are stored on internal disks at the Faculty of Economics and Business handles data storage, data management and data acces in agreement with KU Leuven regulations. Furthermore, data transfer (through Belnet for example) and storage protocols are controlled by the university's Data Protection Officer (DPO).
For particular research tasks in specified work packages, data will be stored on dedicated computing servers to increase computational speeds. In this, an additional layer of security is added on top of the university- and faculty-wide security and access protocols. Access to folders and use of specified data sets is then granted on a case-by-case basis for involved researchers by the system administrators of the respective computing servers.

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution

Data storage, management and acces are handled by the KU Leuven as well as IT-services at the Faculty of Economics and Business (see also above). The latter have the necessary infrastructure and protocols in place to simultaneously backup data on multiple servers.

Is there currently sufficient storage & backup capacity during the project?

If no or insufficient storage or backup capacities are available, explain how this will be taken care of.

- Yes

Storage capacity is provided by (1) the Faculty of Economics and Business through the system of internal disks with allocated slots for separate research groups, (2) disk space on dedicated computing servers, and (3) the KU Leuven through the OneDrive system that provides researchers with both personal and shared storage facilities (should the need arise). All types of storage are sufficient to handle backups of the data used in the project.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

In addition to security measures put in place by the KU Leuven that restrict acces to unauthorised users and guarantee secure data storage, internal procedures are implemented to restrict acces to work package folders and proprietary data to users on a case-by-case basis to prevent improper use of these data sets, see also above.
Furthermore, the research team has established a reputation in data storage and data protection in handling proprietary data sets on a contractual basis with former and current data providers. Proprietary data sets will be stored on the secure FEB data server that has restricted access, or on dedicated computing servers. These data files will be accessible only to researchers named in the contract with the data suppliers. The main risk regarding the safety of personal data is unauthorized access to the files. Physical and software- related measures have been taken by KU Leuven and FEB to protect the servers on which the data are stored. Data files are only accessible after identification. If there has been an event that could potentially be considered a data leak, KU Leuven and FEB follow a protocol to protect the integrity of the data. This process is closely monitored by the privacy-teams supervised by the data protection officer at KU Leuven, see https://admin.kuleuven.be/privacy.
Most of the data sets administered by the Nationale Bank van België (NBB) are confidential. Confidential NBB data sets are stored on internal servers that are disconnected from the outside world and require direct interaction on the secured premises of the NBB. We follow their institutionalized data security standards; these procedures apply to the firm-level transactions and trade data used in work packages 3 and 4.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

Research groups at FEB contribute annually by means of a fixed sum paid to ICT-services to cover basic ICT-support which includes data storage, data backup and server maintenance. The ICT-costs stemming from research activities in this project are budgeted on a proportional basis.

Data Preservation after the end of the Research Project

Which data will be retained for 10 years (or longer, in agreement with other retention policies that are applicable) after the end of the project?

In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

- All data will be preserved for 10 years according to KU Leuven RDM policy

KU Leuven guarantees safe data storage and restricted data access for the duration of the project as well as a minimum period of 10 years after completion of the project. This basically covers publicly available data; storage of proprietary data however is stipulated by contractual obligations and often involve physical removal from disks after completion of the project.

Where will these data be archived (stored and curated for the long-term)?

- Large Volume Storage (longterm for large volumes)

After the ten-year post-completion storage period, project folders and data will be archived on the KU Leuven central servers (including automatic backup procedures) in agreement with KU Leuven's RDM policy.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

Preservation costs will be covered by the annual ICT-contributions discussed above and budgeted on future research projects.

Data Sharing and Reuse

Will the data (or part of the data) be made available for reuse after/during the project?
Please explain per dataset or data type which data will be made available.

- Other (specify below)

Contracts with the suppliers of proprietary data forbid reuse; only aggregate statistics and data can be reused.
Within the bounds of the confidentiality agreements and non-disclosure agreements that rest on using the source data, we will only publish aggregate-level (sector- and/or regional-level) data derived from the unit-level data sources. We will publish information on production and consumption characteristics for specified groups (socio-economic cells and detailed NACE 143 sectors).

If access is restricted, please specify who will be able to access the data and under what conditions.

Each of the data files will only be accessible to the researchers named in the contract with the data suppliers. We will only share all program code with researchers for the purposes of replication. Researchers who obtain their own access to the underlying data sources can then replicate our results. Suitably aggregated data, and accompanying technical notes can be published on the project website.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

Please explain per dataset or data type where appropriate.

- Yes, privacy aspects

Contractual obligations prevent proprietary data to be shared with non-authorised users, or re-used for other purposes. In addition, to ensure privacy and compliance with GDPR regulations, individual-level observations cannot be disclosed or shared with other non-authorised users.

Where will the data be made available?

If already known, please provide a repository per dataset or data type.

- Other (specify below)

Data will be made available though the project website (GitHub); otherwise, KU Leuven RDR will be selected.

When will the data be made available?

- Upon publication of research results

Which data usage licenses are you going to provide?

If none, please explain why.

- Other (specify below)

Not applicable.

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

- No

What are the expected costs for data sharing? How will these costs be covered?

Expected costs for data sharing will be low given the nature (binary files) and size (very small) of the project files that can be posted publicly, and mainly relate to setup and maintenance of corresponding project websites (GitHub). These are also budgeted on the project.

Responsibilities

Who will manage data documentation and metadata during the research project?

Data documentation and metadata are handled by the respective work package leaders, i.e. Prof. Dr. Frank Verboven (WP 1), Prof. Dr. Joep Konings (WP 2), Prof. Dr. Jo Vanbiesebroeck (WP 3), and Prof. Dr. Hylke Vandenbussche (WP 4). Responsibility for coordination and overall management of data documentation and metadata lies with the principal investigator (PI), Prof. Dr. Joep Konings.

Who will manage data storage and backup during the research project?

Data storage and backup are facilitated by means of FEB protocols and managed automatically. Scheduled (monthly) checks of storage and backup for each work package are handled by the respective work package leaders, i.e. Prof. Dr. Frank Verboven (WP 1), Prof. Dr. Joep Konings (WP 2), Prof. Dr. Jo Vanbiesebroeck (WP 3), and Prof. Dr. Hylke Vandenbussche (WP 4). Follow-up and overall checks of project data storage and backup are handled by the PI, Prof. Dr. Joep Konings.

Who will manage data preservation and sharing?

The IT-services at the Faculty of Economics are responsible for the data preservation; data sharing is handled by the respective work package leaders. Prof. Dr. Joep Koning will bear the end responsibility for data preservation and data sharing for the entire project.

Who will update and implement this DMP?

Updates to this DMP and its effective implementation are handled by the PI, Prof. Dr. Joep Konings. He is assisted in his task by the Research Data Management (RDM) team of the KU Leuven, in particular with respect to follow-up of this DMP, see https://www.kuleuven.be/rdm/en for more information.