Acknowledgements and Disclosure

The authors acknowledge financial support from the institutions funding the projects described in this paper. The Micro-Data Infrastructure was developed under the European Union's Horizon 2020 research and innovation programme through **MICROPROD** (grant agreement No. 822390), extended under the European Commission's Technical Support Instrument through **MULTIMSPROD** (DG REFORM grant agreements No. 101101853 and No. 101140673), and is currently supported under the Technical Support Instrument through **ProdTool** (grant agreement No. 101261159). Work on the infrastructure at the Halle Institute for Economic Research (IWH) is additionally supported by institutional funding from the German federal government and the State of Saxony-Anhalt.

Eric J. Bartelsman (Vrije Universiteit Amsterdam), Filippo di Mauro (IWH–CompNet) and Javier Miranda (IWH–Centre for Business and Productivity Dynamics) each acknowledge support from the institutions supporting the grants described above. The authors gratefully acknowledge the collaboration of the national statistical institutes, national productivity boards and secure data access centres participating in the Micro-Data Infrastructure, without whose cooperation none of the work described here would be possible.

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the European Commission, of any participating national statistical institute, or of their respective institutions. The funders had no role in the preparation of this paper or in the decision to submit it.

Beyond the research support disclosed above, the authors have no financial relationships or other potential conflicts of interest to declare relevant to this work.

The Micro-Data Infrastructure (MDI): A Distributed Infrastructure for Harmonised Firm-Level Research Across National Statistical Institutes

Eric J. Bartelsman (Vrije Universiteit Amsterdam) · Filippo di Mauro (IWH–CompNet) · Javier Miranda (IWH & Friedrich-Schiller University Jena)

Submission to the NBER Innovation Information Initiative (I3) Technical Working Group Meeting
4–5 December 2026, Cambridge, MA · Contact: javier.miranda@iwh-halle.de

1. Introduction

The Micro-Data Infrastructure (MDI) is an operational, open-source, R-based, research infrastructure that lets a researcher write a single analysis program and have it executed, unmodified, on confidential firm-level microdata held inside multiple European National Statistical Institutes (NSIs), returning harmonised, disclosure-ready statistical output that can be stacked into a dataset ready for cross-country analysis once released by the NSI. The MDI is currently running or in advanced set-up in eight European countries, is maintained by the Centre for Business and Productivity Dynamics (CBPD) at the Halle Institute for Economic Research (IWH) together with Vrije Universiteit Amsterdam, and is funded through a sequence of three EU grants plus a permanent German federal funding line.

We present the design of the infrastructure, the specific institutional and technical problems it solves, what it can and cannot currently do, how it is documented and taught, how we intend to open it to the broader research community, and how other countries can join the infrastructure.

2. The problem

European business microdata are among the richest in the world and among the least used. Currently, cross-country research and analysis is scarce. National Statistical Institutes across the EU hold high-quality business registers, structural business statistics, innovation surveys, trade transactions, energy use and ICT-adoption data–frequently linkable at the firm level. Yet cross-country empirical work on productivity, firm dynamics, innovation and technology adoption in Europe is still overwhelmingly conducted either on aggregate data or on a single country at a time.[1] There is no European counterpart to the US Federal Statistical Research Data Center system: no single accreditation, no common protocol, no common file layout. This limits access and the types of questions that can be asked of the data.

The obstacles to advancing cross-country policy and research are not primarily about willingness. They are structural:

- **Legal fragmentation.** Access rules, accreditation procedures, permissible output and researcher eligibility differ by country. A project cleared in one jurisdiction confers nothing in the next.
- **Semantic fragmentation.** Even where the same Eurostat regulation governs a collection, national implementations differ in variable definitions, units, sector and product classifications, vintages of those classifications, firm identifiers, and treatment of enterprise groups. Harmonisation done *ex post* on aggregates is lossy and often undocumented.
- **Physical fragmentation.** Microdata cannot be pooled. Any cross-country design must be executed where the data live.
- **Transaction costs.** These are large and highly asymmetric across countries. To make this concrete, the MDI manual records observed turnaround and price for moving code in and results out of partner environments: 1–2 days and €125–250 per output release in the Netherlands; 1–2 days and €318 per ten-unit output package in France; 10–30 days but free of charge in Slovenia; and 21–45 days at roughly €118 per hour in Austria. A design that requires many iterations across seven countries is not only slow; it is unaffordable.

[1] The CompNet aggregated data accessible to researchers upon request, and the confidential OECD aggregated datasets the result of Dynemp, and Multiprod are the source of most current cross-country work, using the distributed micro-data approach developped by Bartelsman et al. [2004] .

The consequence is that the marginal cost of adding the *second* country to a European firm-level study is close to the cost of the first, even without accounting for the researchers time and effort in learning about the data and writing custom code for each country. The MDI is an attempt to drive that marginal cost toward zero.

3. What the MDI is

The MDI is a distributed execution system with a metadata-driven harmonisation layer. It has four components, installed identically at every participating site, including a mock-data development site:

- `rocket` — the harmonisation engine, NSI metadata, auxiliary data and the R tool libraries.
- `payload` — versioned research modules (analysis code) plus launch-specific harmonised-panel metadata.
- `launchpad` — the small set of site-specific control files a local operator edits.
- `docs` — system documentation, including the MDI Users Guide.
- `training` — development, testing, and training server.

A "launch" proceeds in three steps. The local operator edits `countdown.R`, supplying only paths, the ISO country code and three disclosure parameters: the minimum number of contributing firms per cell (`MDIminNumObs`), the dominance share threshold (`MDIdomSh`), and the number of top firms in the dominance rule (`MDIdomNr`). A validation programme, `pre_launch_checker.R`, then audits the national metadata and concordances and writes a diagnostic report before any expensive computation occurs. Finally `liftoff.R` reads the raw national files, builds the harmonised firm-level panels, and runs every research module in the payload, writing pre-disclosure output to a local directory for NSI clearance.

The critical design choice is that harmonisation is expressed as machine readeable metadata, not as country-specific code. Each NSI supplies five machine-readable files describing its own holdings—a list of data files, variable-level metadata, a codebook for categorical variables, classification lists, and time concordances between classification vintages. The MDI side supplies five parallel files defining the target harmonised panels, including classification hierarchies. Two concordance files then map national objects onto the harmonised specification, at the variable level and at the code level. Everything a country does differently is confined to those files. To the analysis code in the payload, every country looks identical. This is what makes a module portable, and it is also what makes the harmonisation auditable: the mapping from national source to harmonised concept is a reviewable artefact rather than an undocumented step inside somebody's script.

3.1 Capabilities

Harmonised panels. The current specification defines twelve firm-level panels, all keyed to the business register as the linking backbone:

Panel	Content	Panel	Content
BR	Business Register (backbone)	ITS	International Trade in Services
BS	Balance Sheet / enterprise groups	OFATS	Outgoing Foreign Affiliates Statistics
BD	Business Dynamics	IFATS	Incoming Foreign Affiliates Statistics
SBS	Structural Business Statistics	ENER	Energy use at firms
CIS	Community Innovation Survey	PRODCOM	Production by firm and product
ICTEC	ICT Usage in Enterprises Survey		
ITGS	International Trade in Goods		

For an innovation-research audience the relevant point is that CIS, ICTEC, PRODCOM and the trade panels are linkable to the register and to balance-sheet data *within* the harmonised environment, so that innovation and technology-adoption measures can be related to firm growth, entry, exit, exporting and productivity on a common firm identifier, in several countries, using one programme.

Tooling. The infrastructure ships an R package organised in five tool families: metadata validation; infrastructure tools for import, harmonisation and disclosure checking; user-facing analysis tools for module writers; reporting and analysis helpers; and programmer utilities. Functions are documented with Roxygen and exposed through standard R help. Disclosure protection is built into the analysis primitives rather than bolted on afterwards: `mdi_aggregate(..., disclosure=TRUE)` evaluates cell counts and dominance against the site's parameters and flags failures, `mdi_regress()` handles the estimation cases, and `mdi_export()` writes releasable output packages. Production-function estimation tools (OLS, LP, ACF, control-function, dynamic GMM and Wooldridge-type estimators) are maintained inside the same codebase, so that productivity measurement is version-controlled alongside the data construction.

Development environment. Module writers develop against confidential data in one country. A cloud test server also allows module writes to develop code against mock data. A cloud server hosts a complete mock-data environment reproducing the full harmonised catalogue with realistic structure but no sensitive content, together with all MDI tools. Code is written, debugged and cost-estimated there before a launch is requested reducing actual time in the confidential environment.

Discovery. A public metadata catalogue viewer lets a prospective user check, before writing anything, whether a variable exists in a given country and for which years—answering the "is this study even feasible?" question without an NDA.

3.2 Member countries and access modes

The MDI adapts to national legal arrangements rather than requiring a single one. Three modes are in use, and the harmonisation layer is identical across all three.

Mode	Countries	Who executes
Remote access	Austria, Germany, France, Netherlands, Slovenia, UK	Accredited MDI staff work inside the NSI environment
MDI partner	Finland, Italy	Partner institution operates the launch locally
Remote execution	Portugal	NSI staff run the submitted rocket and release cleared output

Greece and Malta are being onboarded in the current grant, with Italy and Greece metadata integrated in the launch v2.4 release line. Spain has engaged with the framework through the Bank of Spain. Access at each site rests on standard access protocols available in each country; microdata never leave the NSI.

3.3 Training and documentation

Documentation is treated as infrastructure. The MDI Users Guide covers the NSI-side installation and metadata specification, the module-writer workflow, and launch operations. A three-session training programme, hosted on a dedicated cloud environment that replicates an NSI installation over mock data, takes participants through (i) the launch workflow, disclosure settings and data discovery through the catalogue viewer; (ii) efficient module writing in R with the MDI aggregation, regression and export functions; and (iii) consolidating and visualising output across countries after release. A parallel training track exists for the micro-aggregated CompNet dataset, which serves as the entry point for institutions not yet connected to the MDI. Training is delivered remotely, on-site and in hybrid form.

4. What we are doing with it

Analysis is organised as *research modules*: self-contained, peer-reviewed programmes admitted to a numbered launch. The call for the most recent launch produced a portfolio that illustrates the range the infrastructure supports:

- **Cross-country productivity and firm-level indicators** — state-of-the-art production-function estimates and firm-level moments across sectors (FI, FR, NL, PT, SI).

- **Firm responsiveness over the business cycle** — employment and tangible/intangible capital dynamics, and responsiveness of input choices to productivity shocks, as a measure of allocative efficiency across European countries.
- **High-growth firms and job flows** — an adaptation of the US Census Bureau's BDS-HG methodology to MDI countries, developed and validated on Portugal with a US benchmark, producing harmonised job creation, destruction, and high-growth-firm tabulations by firm age and size.
- **Digital transformation, trade and productivity for MSMEs** — digitalisation gradients from the ICT-usage panel linked to trade participation and productivity.
- **Mapping manufacturing capabilities** — construction of a taxonomy of processes, materials and facilities (NAMTEC) from product-level production and trade data, to characterise national technological capability.
- **Accelerating the green transition** — energy substitution, cost pass-through and carbon-tax effectiveness at the firm level, in cooperation with the OECD.
- **Global spillovers of US tariff policy** — third-market competition effects using firm-level trade transactions.
- **Start-up strategies** — an externally proposed module (EBRD) clustering firms by entry strategy using register, innovation, foreign-ownership and demographic data.

5. Partners, funding and deliverables

The MDI is the cumulative output of three EU grants and now a national funding line.

Programme	Instrument	Period	Contribution to the MDI
MICROPROD	Horizon 2020, GA 822390, €2.98m	2019–2022	Origin of the MDI, extending the Micro Data Linking project led by Statistics Denmark
MULTIMSPROD	DG REFORM / TSI, GA 101101853 and 101140673	2022–2025	Country expansion (PT, SI, DE, AT), NSI agreements, first NPB trainings, CompNet Data Access Toolkit
ProdTool	TSI, GA 101261159, €1.52m	Dec 2025–May 2028	Consolidation, onboarding of IT/MT/GR, proposal-approval process, dashboards, opening to external users

The current consortium is coordinated by IWH–CBPD, with Vrije Universiteit Amsterdam and Bocconi University; beneficiaries are the National Productivity Boards of Germany, Austria, Slovenia, Finland, the Netherlands, Italy, Malta and Greece, with France and Portugal participating through existing MDI arrangements and Spain soon to join. Funded deliverables include country MDI set-up and metadata completion, streamlining of the launch process, yearly data updates, expansion of the harmonised catalogue, MDI and CompNet visualisation tools, the training programme, an annual jointly authored Common Report with published replication code, and—explicitly—a deliverable on opening the MDI to users beyond the consortium. Separately, a permanent federal funding line has been awarded to IWH, which is what makes a standing operational unit possible rather than a rolling sequence of project hires. Recent and forthcoming outlets include the ProdTool Common Report 2026 on Europe's productivity challenge, the first ProdTool workshop (Vienna, February 2026) and the second (Brussels, October 2026) alongside the 15th Annual IWH–CompNet Conference at the European Commission.

The MDI's sibling product, the CompNet dataset, is relevant context: micro-aggregated firm-level indicators and distributional moments for 18 European countries covering 1995–2023, including Olley–Pakes and Foster-type decompositions, now in its tenth vintage. CompNet demonstrates the demand; the MDI is the response to CompNet's limitation, namely that a fixed set of pre-computed aggregates cannot answer a question nobody anticipated.

6. Opening the MDI to the research community

The MDI is operational. It is installed across member environments, documented and taught, and it supports a growing pipeline of research modules. We have demonstrated that the model works: external code can be written against a harmonised catalogue and executed on confidential national microdata, with cleared results returned to the researcher. The next step is to open it progressively to a broader research community, and we are building scalable arrangements to make that possible.

6.1 The model works

France is the working example. Two agreements with the Centre d'accès sécurisé aux données (CASD) establish it: an infrastructure agreement under which IWH–CBPD provides and maintains the MDI inside CASD's secure environment, and a replication agreement governing execution of externally developed code on French microdata.

What makes this portable is the infrastructure itself. Because modules are written against a harmonised catalogue and a fixed set of tool signatures rather than the idiosyncrasies of one national holding, the same script runs in another member environment; and because disclosure parameters are enforced inside the tools, output arrives at the clearance officer already close to releasable. The two-contract structure is a template, designed to be repeated with each national provider rather than renegotiated from scratch.

6.2 Entry is cheap—for researchers and for institutes

1. **Tooling as an installable package.** The infrastructure and analysis layer is packaged as a public R package, so hosting it is a versioned software installation rather than a bespoke integration. An institute installs the package and supplies its national mappings and disclosure parameters; it maintains only the part no one else can.
2. **Mock and synthetic data.** A mock-data test server already lets developers run the full launch sequence end to end with no confidential records, and synthetic data with realistic joint structure is in production. A researcher can see what variables exist and write a complete module outside the vault that runs correctly on first execution inside it.
3. **Built-in disclosure.** Disclosure rules are enforced inside the tools and parameterised per institute rather than applied by hand, so clearance becomes verification rather than reconstruction.

6.3 What opening up delivers

Standardised execution against a harmonised catalogue with versioned tools means code, software versions and source vintages are documentable and shareable without releasing a single confidential record.

- *Computational reproducibility.* Re-executing the same module on the same vintage returns the same released numbers, verifiable inside a trusted environment without the verifier ever holding the data.
- *Replication and generalisability.* The identical module runs on independent national datasets, turning cross-country testing into a by-product of the design.
- *Cumulative knowledge.* Estimators, concordances and decompositions become reusable, versioned code rather than project-specific work. Knowledge builds on knowledge because a methodological advance is a piece of code that can simply be passed on, and a public module repository makes those contributions discoverable and citable.

6.4 Putting it on a permanent footing

A permanent German federal funding line has been awarded to IWH, which converts the MDI from a project deliverable into a standing service with a team that does not disappear when a grant ends. Beyond that core, we expect the mature model to combine continued research funding, institutional support from participating members and, as is already standard practice at national access centres, a contribution from users—a mix that is well established in this field and that makes progressive opening realistic.

6.5 An invitation

We welcome feedback from users and providers. Specifically we welcome feedback on the following elements. First, **pilot users**: researchers with a well-defined cross-country question who would work with us as the first external projects, and whose experience would shape the access process we are putting in place. Second, **partners**: national providers and research infrastructures interested in hosting an MDI node, for whom the CASD template is available and the installation cost is low. Third, the **I3 community's advice** on a question we cannot settle alone. Because the data cannot be deposited, we can instead publish a complete record of how a result was produced—module version, tool versions, data vintage, disclosure settings—and have its execution certified by a trusted third party under replication agreements at each node, in the mode of what we already do in France. What would such a certification need to contain for journals and archives to accept it in place of a data deposit? That is a standard the research community has to set, and we would like to help shape it.

References

Eric Bartelsman, John C. Haltiwanger, and Stefano Scarpetta. Microeconomic evidence of creative destruction in industrial and developing countries. IZA Discussion Paper 1374, Institute for the Study of Labor (IZA), Bonn, October 2004. URL `https://docs.iza.org/dp1374.pdf`.