IEP@BU Working Paper Series

Bocconi

IEP-COMPET Dataset

Daniel **Gros**
Silvan Michael **Hofer**
Philipp-Leo **Mengel**
Mauro **Molteni**
Giorgio **Presidente**
Cristina **Rujan**
Florian **Schimmel**

Università
Bocconi
IEP@BU
Institute for European
Policymaking

June 2025

IEP-COMPET Dataset

Daniel Gros[1], Silvan Michael Hofer[1], Philipp-Leo Mengel[1], Mauro Molteni[1], Giorgio Presidente[1], Cristina Rujan[2], and Florian Schimmel[1]

[1]*Institute for European Policymaking, Bocconi University*
[2]*ifo Institute – Leibniz Institute for Economic Research at the University of Munich*

June 18, 2025

Abstract

This paper presents the construction of the IEP-COMPET dataset, which links approximately two-thirds of all EU Horizon 2020 and Horizon Europe grants awarded to companies with their financial performance and patents, using various data sources. The dataset can be used to examine the allocation and impact of EU funding for innovation. Interested researchers should contact giorgio.presidente@unibocconi.it

Keywords: C80,L53

1 Introduction

This paper documents the construction of the IEP-COMPET dataset, a panel dataset linking EU-funded R&D and innovation support measures to firms' financials and patents.

2 Overview of EU Funding

The European Union funds R&D and innovation through what are known as *framework programs*, multi-annual budget allocations set every seven years by EU officials. The EU has had multiple Framework Programs since the early 1980s. The most recent (and largest) two are:[1]

- Horizon 2020 (2014–2020)
- Horizon Europe (2021–2027)

Within each Framework Program, there are many different *funding instruments*, each with its own characteristics and participation requirements.[2]

Entities can request support by applying to *funding calls* on the online "EU Funding & Tenders Portal", which publishes calls issued under the different funding instruments.

The two main sources of the IEP-COMPET dataset are the European Commission's CORDIS (Community Research and Development Information Service) and Bureau van Dijk's ORBIS.

Further funding and firm-level characteristics are collected by Crunchbase.

3 CORDIS Files

We use two files from CORDIS, which are described below.

3.1 organizations.dta

This file is taken from this link: `https://data.europa.eu/data/datasets/cordisref-data?locale=en`

The file provides information by "organization id", a unique identifier of the funding beneficiary.

[1] We do not consider the previous, FP7 (2007–2013)–The Seventh Framework Program

[2] One example is the "SME Instrument", an accelerator for start-ups and SMEs more broadly, active under Horizon 2020. Another example is "Pathfinder", an instrument focusing on projects still far from commercialization. The latter belongs to the institutional umbrella of the "European Innovation Council", which is itself part of Horizon Europe.

Each organization can receive multiple funding from the same instrument and/or participate in multiple funding instruments/calls. Therefore, the same id can appear multiple times in the data.

At the same time, for each organization id, the file lists the projects to which it participates, each uniquely identified by a project identifier (see below).

The file provides the EUR amount received by an organization. Since most funding programs involve collaborations across entities, the amounts reported in each entry of the file are not necessarily equal to the total project amount awarded by a funding call. However, the sum across all consortium members is equal to the total project amount.

Most organizations in the file receive some funding. However, some funding calls involve multiple-entity-consortia which include "third parties", i.e. entities that do not receive financial support.

Organizations can be of 5 types: Private Company (PRC), Higher Education (HES), Research Organisation (REC), Public Body (PUB) & Other entities (OTH), which comprise mostly NGOs.

3.2 project.dta

This file is taken from this link: https://data.europa.eu/data/datasets/cordisref-data?locale=en

The file provides information by "project id", which uniquely identifies a project funded by the EU within a specific funding call.

For each project, it further includes the code of the call, the name of the funding instrument through which it is awarded, and the respective framework program.

The file also provides the signing date (co-signing if involving multiple participants), start date, and end date of the projects.

3.3 CORDIS Files Merged

A one-to-many merge of project.dta to organizations.dta allows to assign time stamps for the funding to organizations corresponding to calendar dates. As noted, three dates are considered: signing, start, and end date.

The information is aggregated at the year level. If an organization receives multiple grants from the same funding instrument within a year—for instance due to participation in different calls—the amounts are simply summed at the year level.

If an organization receives multiple grants from different funding instruments within a year, the amounts are summed by funding instrument and year.

Thus, the merged CORDIS dataset vary at the organization-funding instrument-year level.

4 ORBIS File

For every "organization id" in CORDIS (n=32,829), we find the corresponding organization in the ORBIS database[3] from Bureau van Dijk (2025) and assign its "BvD Number". ORBIS is a database that provides a range of company information. We use it for retrieving financial data, like turnover and employment.

To match CORDIS to ORBIS we proceed as follows: When available, we use the European VAT number from CORDIS to find a direct match in ORBIS (n=23,340). If unavailable, we use ORBIS' "batch-search" tool to look for suitable matches based on name, city, and country (n=9,899). If we do not find a company manually, we classify it as not found (n=1,071)[4].

For the VAT numbers, ORBIS returned all companies with the given VAT number. This included not only the Global Ultimate Owner (GUO) but also branches, subsidiaries, single locations, and independent cooperations. Overall, we get 163,285 potential matches. We use the country, city, postcode, street, and street number to identify the correct and final matches. This exercise is approached from two angles. First, we focus on finding unique matches using combinations of these data points. Second, we exclude firms that clearly do not match.

In the first step, we look for unique matches as combinations of the available information. Using different combinations reduces the sensitivity to misspellings. For example, if we were to match all data points, we might not find accurate matches due to inconsistent spelling (e.g., Munchen, Muenchen, or München). By focusing on subsets of the information (e.g., the same country, postcode, and street), we can confidently identify the correct match, even if not all data points align. Importantly, we only consider a match valid if it is unique. For example, if two firms match the description due to different street numbers, we would not make a decision and would

[3] https://login.bvdinfo.com/R1/Orbis

[4] Careful readers will spot that $32{,}829 \neq 23{,}340 + 9{,}899$. That is because 410 companies with VAT codes were also matched manually. If available, the manual matches were chosen.

leave the match unresolved. Using this approach, we identify 20,703 of the 23,340 firms. For the remaining firms, there are still 41,418 potential matches to consider.

In the second step, we remove implausible matches within the remaining subset of 41,418 observations. For example, if we have 10 matches for one VAT number, but only half of them are in the right country, we remove those matches accordingly.

After these two steps, we are left with 29,772 matches. To account for the ambiguity, we include a weight-variable to the dataset that indicates the number of remaining matches (i.e. it would be equal to five in the previous country example), and potential matches (i.e. ten).

In the manual search, we upload the name, city, and country to ORBIS' batch-search tool. ORBIS then returns candidate matches with a confidence score from A to E. We take matches with score A as correct. For all others, we check the candidates and choose the most plausible one. If none of the candidates is plausible, we classify as no match. This is the case for 1,071 out of 32,829 companies. Because of some anomalies, we also match a small subset of the VAT companies manually but remove the VAT duplicates from the final ORBIS dataset.

Finally, we combine the data from the two matching procedures to obtain a dataset with 38,505 observations, 8,828 from the manual process, 28,606 from the VAT matching (22,930 firms with multiple matches), and 1,071 which could not be found.

After linking ORBIS to CORDIS, we proceed to downloading the data . We download ORBIS information for all available firm, i.e. 38,505 - 1,071 = 37,434. Of these, 37,357 firms were unique, as some firms were matched to multiple CORDIS firms.[5] Of those, only one company could not be found for the download, even when using the BvD ID.

To keep all CORDIS firms in the data frame, we left-join the ORBIS data to CORDIS observations. So, the data comprises 38,505 observations again—including all CORDIS IDs, with some otherwise empty rows.

The final ORBIS dataset contains 38,505 observations.

[5] Looking at CORDIS, we suspect some companies in CORDIS to be duplicates

5 Seal of Excellence

We follow the same steps outlined above to obtain ORBIS information for companies awarded the Seal of Excellence (SoE). The SoE is a quality label given to high-ranking Horizon Europe project proposals that are not funded due to budget constraints—hence these companies are not necessarily included in CORDIS. [6].

We retrieved the data for the SoE firms from Dealroom (Demolin et al., 2025). They compiled a non-exhaustive list of 814 firms that have received the Seal.

We match them to CORDIS firms using first the Levenshtein distance to get potential matches, and confirm them manually. Through this process we identify a subset of the SoE companies are also listed in CORDIS (213 firms, indicated via a dummy). This is because these firms previously applied for an EU grant unsuccessfully but later received EU funding in a subsequent funding call. For the other Seal of Excellence companies we perform another manual batch search in ORBIS and download the batch (518 firms). We add both batches to the rest of the ORBIS data and tag it with a corresponding dummy variable. The remaining firms could not be found and are not added to any dataset.

6 Orbis IP file

We rely on data from the Orbis IP database [7] to build patent-based innovation outcomes. That is, we retrieve all patent publications owned by the firms of interest, given the BvD Numbers identified, as suggested above. Out of the 37,357 linked firms, 12,637 (34%) own at least one patent. Overall, roughly 10 mil. distinct patent publications are associated with these firms.

To avoid counting the same invention multiple times, we rely on *patent families* instead of individual patent filings, or corresponding publications. The ORBIS IP database provides indicators for patent families, including simple patent families, which are equivalent to DOCDB patent families used in PATSTAT, the patent database of the European Patent Office. Such a patent family comprises a set of patent filings across jurisdictions, protecting the same underlying invention or technology. For brevity, we refer to a patent family simply as a patent in the report. We date each patent by its priority date, i.e., the date of the earliest filing within the

[6] https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/seal-excellence_en

[7] Data extraction: June 2024.

family which is the closest in time to the origin of the invention.

We assign each patent family a main technology class, by identifying its main International Patent Classification (IPC) subclass (4-digit code), i.e. the subclass that is most common across all filings in the patent family. Whenever a mode cannot be identified, we randomly choose one subclass from those that are most common. In a second step, based on the subclass identified, we assign each patent family a main technological area based on the classification introduced by Schmoch (2008).

We follow the methodology provided by Eurostat (2024) to classify high-tech patents. Given the different granularity level of technology subclasses considered in this classification, we rely on the complete set of IPC codes assigned to each patent in the family. A patent family is classified as high-tech if at least one of these IPC codes matches a code listed in Eurostat's high-tech classification.

When aggregating patents at the firm-year level, we assign a co-owned patent to each of the awardees that acts as an owner.

7 Crunchbase File

Crunchbase is a widely used database that provides comprehensive information on both private and public companies, including details about founders, key executives, investors, and funding rounds (Crunchbase, 2025). Established in 2007, the database has significantly expanded in scope and coverage over the years.

The Crunchbase dataset is structured into three primary categories: organizations, people, and investments. "Organizations" includes data on companies; "people" captures information on individuals affiliated with these organizations, particularly those in leadership or key decision-making positions; "investments" encompasses data on funding rounds. It also provides granular details such as funding event dates, amounts raised, and investment types (e.g., Seed Funding, Angel Funding, Series A, B, C, etc.). A detailed glossary of funding types is available online at https://support.crunchbase.com/hc/en-us/articles/115010458467-Glossary-of-Funding-Types.

To ensure data accuracy and completeness, Crunchbase employs artificial intelligence and machine learning algorithms for validation and anomaly detection, supplemented by manual curation from a team of data analysts.

For our purposes, we focus on the funding dataset, which includes details as the number of funding rounds, the date of the announced funding, amounts raised, and investment types.

The dataset used in this study was collected between January and February 2025. For each “organization id” in the CORDIS database, we identified the corresponding entity in Crunchbase and retrieved the relevant funding information. The data extraction process followed these steps:

1. Entity Matching via API-Search: Given an entity name in CORDIS, we queried Crunchbase’s Autocomplete API,[8] which returns a list of suggested entities based on the input string. The output includes a list of permalinks corresponding to potential matches.

2. Retrieving Organization Details: For each permalink obtained in the previous step, we retrieved the corresponding Crunchbase location information using the "Lookup an Organization" API.[9]

3. Entity Disambiguation Using Text Similarity: To refine entity matching, we employed a similarity score between CORDIS Entity Name and the suggested Crunchbase Entity Names. Using the Python library "linktransformer",[10] we generated text embeddings and computed a similarity score based on vector space distance. The transformer model "all-MiniLM-L6-v2" from Hugging Face[11] was used due to its efficiency and domain suitability.

4. Final Entity Selection and Manual Review: Using the similarity score and location information, we identified the best-matching Crunchbase entity for each CORDIS organization. A manual review was conducted to ensure accuracy.

5. Retrieving Funding Information: Once a single permalink was assigned to each CORDIS entity, we extracted Crunchbase’s funding information, finalizing a dataset that included: "Crunchbase permalink", "Announced Date", "Transaction Name", "Number of Investors", "Money Raised", "Lead Investors". The "Money Raised" column was parsed to extract the currency and funding amounts, which were then converted to EUR for consistency.

7.1 Crunchbase File Summary Statistics

The dataset comprises 11,139 Crunchbase permalinks, from which we retrieved funding information. Among these, 4,790 permalinks contain details on Money Raised and Announced Date,

[8] https://data.crunchbase.com/docs/using-autocomplete-api

[9] https://data.crunchbase.com/reference/get_data-entities-organizations-entity-id-2

[10] https://github.com/dell-research-harvard/linktransformer

[11] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

covering a total of 10,494 funding events. The funding amounts are reported in multiple currencies, with 57.80% in EUR and 22.83% in USD. The earliest recorded funding event dates back to 1996, while the most recent one is from February 2025. Notably, 80% of funding events were announced between 2015 and 2023.

On average, companies in the dataset received 2.5 funding rounds, with a median of 2 rounds per company. The number of fundings per company ranges from a minimum of one to a maximum of 21 rounds. Regarding funding types, 80% of all funding events fall into four main categories: Grants (36.7%), Seed Funding (18.5%), Venture Rounds (16.9%), and Series A & B (13.1%).

8 Assembling IEP-COMPET

To assemble IEP-COMPET, we perform a 1:m merge of the ORBIS file with the CORDIS merged file. Merging the funding information with the firms' financial data requires setting the 'treatment' year. Crucially, the CORDIS dataset includes the start dates of funded projects but does not specify the exact date when a firm received the subsidy. However, as noted by Mulier and Samarin (2021), most firms typically received the subsidy within eight months of submitting their application. Accordingly we use the project start date as a good approximation to the 'treatment' year.

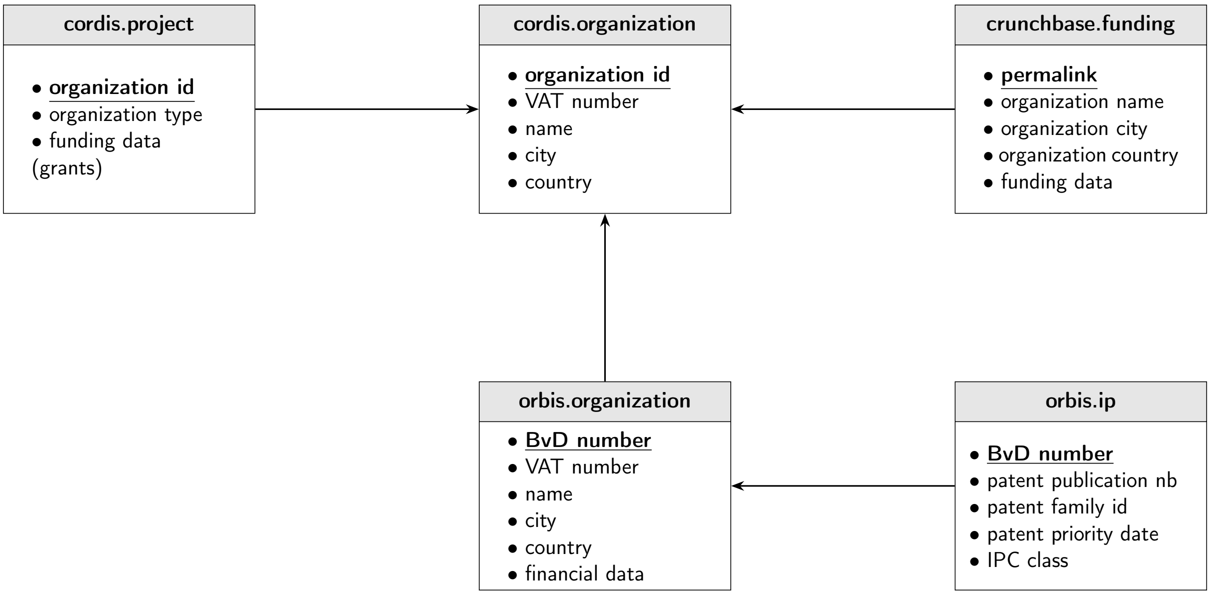
We are able to match two-thirds of the universe of grants disbursed by the Commission. Figure 1 presents graphically the various components of the dataset.

8.1 Data Cleaning

We clean the financial data from ORBIS following a procedure similar to Kalemli-Özcan et al. (2024). Specifically:

- Keep company-years with non-missing information on sales, employment and total assets.
- Drop the entire company if sales, employment or total assets are negative in any year.
- Drop companies with employment larger than Walmart (2 million) in any year.
- Winsorize all balance-sheet and CORDIS variables below 1st & above 99th percentile.

Figure 1: Relationship between data sources



9 Summary Statistics

IEP-COMPET includes information on 23,574 firms operating in 20 one-digit NACE industries between 1995 and 2024. The dataset includes 17,675 funded projects and 40,081 individual grants to companies disbursed along two framework programs and 619 funding instruments.

Table 1 presents key summary statistics of IEP-COMPET.

Table 1: Summary statistics

	(1) N	(2) Mean	(3) Median	(4) SD	(5) Min	(6) Max
Operating Revenue/Turnover (billion €)	150,230	0.593	0.001	2.621	0	20.63
Number Employees	162,448	1,374	30	6,284	0	51,114
# Patents	59,985	30.11	3.00	217.18	1	12,258
# High-tech Patents	59,985	6.24	0.00	84.84	0	6,205
Company EU grant (million €)	38,261	0.367	0.254	0.358	<0.001	1.409
Company funding per year (million €)	38,261	0.119	0.084	0.125	<0.001	1.619
Grant per Employee (million €)	30,739	0.0337	0.001	0.0892	<0.001	1.409
Duration of the project (years)	46,323	3.454	3.500	1.221	0.331	6.245
Size of the consortium	46,323	19.95	15.5	21.38	1	207
Share of research institutions	46,323	0.348	0.340	0.215	0	1

Notes: This table contains summary statistics for the grant recipients included in CORDIS. The annual patent counts are computed for the period 2005-2022 and include only firm-year observations with at least one patent.

References

Bureau van Dijk (2025). Orbis database. https://www.bvdinfo.com/en-us/our-products/data/international/orbis. Accessed: 2024-12-01.

Crunchbase (2025). Crunchbase. https://www.crunchbase.com. Accessed: 2024-12-01.

Demolin, M., T. Povel, and D. Germano (2025). Seals of excellence. In *Dealroom.co*.

Eurostat (2024). High-tech Industry and Knowledge-Intensive Services Indicators. Annex 6: High-tech Aggregation by Patents. https://ec.europa.eu/eurostat/cache/metadata/Annexes/htec_esms_an_6.pdf. Accessed: 2025-05-19.

Kalemli-Özcan, Ş., B. E. Sørensen, C. Villegas-Sanchez, V. Volosovych, and S. Yeşiltaş (2024). How to construct nationally representative firm-level data from the orbis global database: New facts on smes and aggregate implications for industry concentration. *American Economic Journal: Macroeconomics 16*(2), 353–374.

Mulier, K. and I. Samarin (2021). Sector heterogeneity and dynamic effects of innovation subsidies: Evidence from horizon 2020. *Research Policy 50*.

Schmoch, U. (2008). Concept of a technology classification for country comparisons. Final Report to the World Intellectual Property Organisation, WIPO.