European Commission

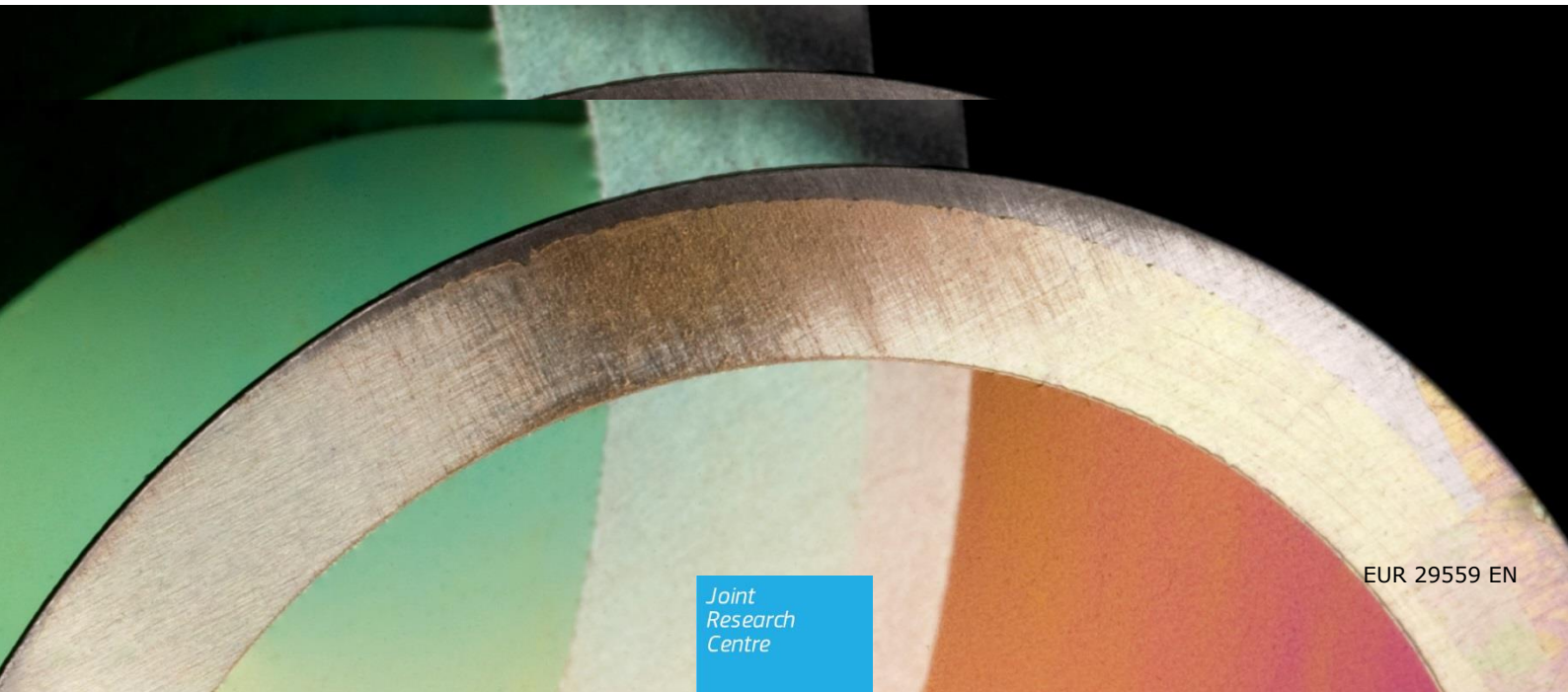
JRC TECHNICAL REPORTS

Advancing the Innovation Radar

Enhancing Innovation Radar data with financial, patent and Venture Capital data

Vincent Van Roy, Tom Magerman and Daniel Nepelski

2018



Joint Research Centre

EUR 29559 EN

EU Science Hub
https://ec.europa.eu/jrc

JRC114418

EUR 29559 EN

PDF ISBN 978-92-79-98370-2 ISSN 1831-9424 doi:10.2760/127887

Luxembourg: Publications Office of the European Union, 2018

How to cite: Van Roy, V., Magerman, T. and Nepelski, D., *Advancing the Innovation Radar. Enhancing Innovation Radar data with financial, patent and Venture Capital data*, EUR 29559 EN, Publications Office of the European Union, Luxembourg, 2018, ISBN 978-92-79-98370-2, doi:10.2760/127887, JRC114418.

Title Advancing the Innovation Radar. Enhancing Innovation Radar data with financial, patent and Venture Capital data

Abstract
The Innovation Radar (IR) is a European Commission (EC) initiative to identify high-potential innovations and innovators in EC-funded Framework Programme (FP) research and innovation projects and to guide project consortia in terms of the appropriate steps to reach the market. This report presents the process and results of linking the IR data with third-party databases to obtain performance information about the innovators in FP projects identified by the IR. In particular, IR participants identified between March 2014 and January 2018 are enriched with financial information from ORBIS, patent information from PATSTAT and private funding information from Dealroom. This enriched data warehouse aims to facilitate the profiling of IR participants in terms of performance, which can subsequently provide guidance for hands-on policy support initiatives. It creates foundation for future analysis of the determinants and barriers of innovation in EU-funded collaborative projects.

Table of contents

Foreword

This report is prepared in the context of the three-year research project on Research on Innovation, Start-up Europe and Standardisation (RISES), jointly launched in 2017 by JRC and DG CONNECT of the European Commission. The JRC provides evidence-based support to policies in the domain of digital innovation and start-ups. In particular:

- Innovation with the focus on maximising the innovation output of EC funded research projects, notably building on the Innovation Radar;
- Start-ups and scale-ups – providing support to Start-up Europe; and
- Standardisation and IPR policy aims under the Digital Single Market priorities.

This research builds on the work and expertise gathered within the EURIPIDIS project.

In this report we present the process and results of linking the Innovation Radar (IR) data with third-party databases to obtain performance information about the organisations that participated in FP projects screened by the Innovation Radar. In particular, organisations in FP projects screened by the Innovation Radar between March 2014 and January 2018 are enriched with financial information from ORBIS, patent information from PATSTAT and Venture Capital funding information from Dealroom. This enriched data warehouse aims to facilitate the profilation of IR participants in terms of performance, which can subsequently provide guidance for hands-on policy support initiatives.

Executive summary

The European Commission's (EC) Framework Programme (FP) constitutes an important share in R&D expenditures in Europe. Many EC-funded research projects produce cutting-edge technologies. However, there is a feeling that not all of them reach the market. The question is why? Launched in 2014, **the Innovation Radar is a joint DG CNECT-JRC initiative to identify high-potential innovations and innovators in EC-funded research projects** and guide project consortia in terms of the appropriate steps to reach the market. Its objective is to maximise the outcomes of public money spent on research. Following its successful launch, the Innovation Radar is becoming the main source of actionable intelligence on innovation in publically-funded research projects in Europe. Enrichment of the Innovation Radar database with other data sources is primordial to increase the quality and depth of the intelligence that can be extracted about innovators and innovations in EU-funded projects.

This report presents the process and results of linking the Innovation Radar data with third-party databases to obtain performance information about the organisations that participated in FP projects screened by the Innovation Radar. More concretely, the 5301 organisations in FP projects screened by the Innovation Radar between March 2014 and January 2018 are enriched with:

- Financial information from ORBIS;
- Patent information from PATSTAT;
- Venture Capital (VC) funding information from Dealroom.

Table 1 presents an overview of the matching results of Innovation Radar with ORBIS, PATSTAT and Dealroom databases and highlights the numbers and percentages of matched observations by organisation types.

Table 1: Matching results of Innovation Radar with third-party databases

Organisation type	Total	Financial information		Patent information		VC funding information *	
Universities	690	124	18.0%	580	84.1%	-	-
Research Centers	671	196	29.2%	368	54.8%	-	-
Large firms	1322	1218	92.1%	624	47.2%	72	11.8%
SMEs	1871	1758	94.0%	558	29.8%	119	11.8%
Governmental institutions	412	41	10.0%	65	15.8%	-	-
Others	335	68	20.3%	19	5.7%	-	-
Total	**5,301**	**2,976**	**56.1%**	**2,214**	**41.8%**	**191**	**11.8%**

Note: The table presents an overview of the final matching result of organisations of the Innovation Radar database with financial, patent and VC funding information. Matching results are presented by organisation type. Financial, patent and funding information are respectively based on the ORBIS, PATSTAT and Dealroom databases. *: The VC funding information has only been retrieved for private organisations identified by the IR as key innovators in FP projects.

Calculations: JRC

1 Introduction

The Innovation Radar (IR) is an initiative supported by the European Commission focussing on the identification of high potential innovations and the key innovators behind them in FP7, CIP and Horizon2020 projects with an ICT theme (De Prato et al., 2015). The IR serves as a monitoring tool for policy makers and project officers at the European Commission as it provides up-to-date information on the innovative output of these projects.

Data of the Innovation Radar stem from a questionnaire co-developed by DG CONNECT and DG JRC. The questionnaire is conducted by external experts commissioned by DG CONNECT during periodic reviews of the research projects. Between March 2014 and January 2018 the Innovation Radar monitored the ICT research actions and the e-infrastructures activity under the seventh Framework Programme 2007-2013 (under cooperation and capacities themes), the policy support actions carried out under the competitiveness and innovation framework policy support programme (CIP ICT PSP) and the ICT-related projects in Horizon 2020 (EC, 2014). Since 2018, the Innovation Radar Survey was gradually scaled up to FP projects with other thematic themes to eventually be incorporated as a standard policy tool in the 9th Framework Programme.

Table 2: Overview of innovation projects and organisation types in the Innovation Radar

Review period	May 2014 - January 2018
Number of reviewed projects	1115
Number of innovations	2915
Review type	
First	29.0%
Interim	31.0%
Final	40.0%

Number of unique organisations	All organisations	Key innovators
Total	5301	2037
SMEs	35.3%	39.1%
Large firms	24.9%	24.2%
Universities	13.0%	18.6%
Research Centers	12.7%	12.7%
Government/Others	14.1%	5.4%

Note: Data source - based on the Innovation Radar and CORDIS.
Calculations: JRC

Table 2 provides an overview of the sample of innovation projects and organisation types that have been screened by the Innovation Radar between March 2014 and January 2018. During its pilot phase, the IR survey has been administered to 1115 FP projects. As a result, 2915 innovations were identified. This means that, on average, every project produced between 2 and 3 innovations. Projects are reviewed three times during the project duration. The number of unique organisations active in these projects amounted to 5301. We distinguished six types of organisations, including universities, research centres, small- and medium-sized enterprises, large firms, governmental institutions and others. SMEs and large firms constitute the largest part of key innovators with respective shares of 35% and 25%. Universities, research centers, governmental institutions and other types of organisations account for roughly 13-14 percent each. From all unique organisations, the Innovation Radar identifies 2037 of them as key innovators behind the

innovations developed in the FP projects screened by the Innovation Radar. This corresponds to roughly 38 percent of all organisations.

Enrichment of the Innovation Radar database with other data sources is primordial to increase the quality and depth of the intelligence that can be extracted about innovators and innovations in EU-funded projects. **This report presents the linkage of the Innovation Radar data with third-party databases to obtain more detailed information about the organisations that participated in FP projects screened by the Innovation Radar.** More concretely, the 5301 organisations in FP projects screened by the Innovation Radar as presented above are enriched with:

- Financial information from ORBIS;
- Patent information from PATSTAT;
- External funding information from Dealroom.

The remainder of this report is structured as follows. Section 2 presents the matching procedure between Innovation Radar and ORBIS which will allow the assessment of the economic performance of FP organisations based on company level financial information. Section 3 describes the process and results of the linkage between the Innovation Radar and the EPO PATSTAT Worldwide Patent Statistical Database. Section 4 details the procedure to enrich Innovation Radar data with Venture Capital financing information from the Dealroom database. Finally, section 5 provides concluding remarks and lessons learned about the matching process of the Innovation Radar data with third-party databases.

2 Financial information from ORBIS

This chapter presents the matching procedure and results of linking organisations of FP projects that have been screened by the Innovation Radar (IR) with their financial and productivity activities. Financial and productivity activities of IR participants are retrieved from ORBIS, which is commercialised by Bureau van Dijk Electronic Publishing (BvD). The ORBIS database provides data on firms' financial and productive activities from balance sheets and income statements together with detailed information on firms' domestic and international ownership structure for over 300 million companies across the world.

The outline and information included in this chapter builds extensively on a paper of Kalemli-Ozcan et al. (2015b) that describes how to construct nationally representative firm-level data from the ORBIS global database. In addition, it benefits hugely from a very comprehensive overview of Kalemli-Ozcan et al. (2015a) reviewing the structure of the ORBIS database and providing useful information on how to process ORBIS data.

2.1 Access to ORBIS

Access to the ORBIS databases can be obtained in different ways:

- Online access through an end-user interface (online platform);
- Disk access through BvD historical disks.

Both options have their advantages and drawbacks. The first option is more user-friendly and can be licensed at a lower budget. Being updated regularly, the online platform contains the most up to date version of the ORBIS database, while researchers using disk access gets only updates twice per year (in September and March). However, this advantage needs to be qualified as on average the ORBIS database has a reporting lag of roughly 2 years (Kalemli-Ozcan et al., 2015b).

Despite these advantages, online access may suffer from slow download speeds when extracting large amounts of data. Moreover, the online access does not allow to trace firms back in time as long as the disk access does. Hence, **to maximise the historical time-series of financial information that can be downloaded, JRC favoured a licence through disk access** as this allows for longitudinal firm-level analyses.

ORBIS March 2018 version is used for this matching project. The ORBIS database is not freely accessible but requires an annual subscription. For more information about subscription options we refer to the ORBIS website.

2.2 ORBIS data

The ORBIS database is split into two modules providing different type of firm-level information: 1) the financial module containing firm financial information and 2) the ownership module describing linkages between a subsidiary and its parents and hence allowing to retrieve the complete ownership structure of a firm.

A unique ID number is assigned to each firm to facilitate data collection across the different modules. This unique ID number is also used in other databases owned and commercialised by BvD (e.g. Zephyr database that tracks Merger and Acquisitions activities). Hence, these unique ID numbers – commonly denoted as BvD ID numbers – allow to load the same company file from any other BvD application and database.

BvD ID numbers have a length that varies from 3 to 18 digits and from which the first two characters correspond to the two digit ISO country code where the company is incorporated. The rest of a BvD ID number may vary, according to the following rules[1]:

- When a national ID is available, unique and stable, BvD will favor that ID to build the BvD ID number. It may be a tax number, a trade register number, any type of national ID that identifies a firm;
- When the national ID is not available or available for less than 70% of the firms of a country or not stable or unique, BvD will prefer using the internal identifier of the information provider;
- When none of the above are available, BvD will create its own internal ID.

As **BvD ID numbers are designed as primary keys of BvD databases**, they should never or very rarely change. However, in practice this is not always the case. BvD ID numbers may change when the national ID numbers change in the official data sources. By consequence, different releases of the historical disks may contain different BvD ID numbers that identifies same firms. In order to keep track of these changes, BvD keeps concordance tables and has developed a BvD ID Change Lookup tool that can be accessed at http://idchanges.bvdinfo.com/ (Kalemli-Ozcan et al., 2015b).

2.2.1 Financial information

Data in the financial module in ORBIS primarily stem from firms' financial statements and contain **time-series of various balance sheet items, profit and loss account items, and financial ratios**. Historical disks of ORBIS provide the same underlying financial information in three different versions, differing in reporting currency. Financial information is provided for the following currency types:

- Original currency in which the companies file financial information;
- US dollar currency, applying time-varying exchange rates;
- Euro currency, applying time-varying exchange rates.

To facilitate cross-country comparisons, all the data presented in this report has been retrieved in Euro currency.

In terms of country representation, historical disks of ORBIS cover firm financial information for roughly 200 countries. There is a large variation in country coverage, though European countries tend to have a better coverage.

In terms of time representation, ORBIS has on average a time-lag of roughly 2 years. Hence, at the moment of writing this report, most time-series data ends in 2016 with limited observations for the period 2017-2018. Earliest observations in the historical ORBIS disks date back to the 1970's. Kalemli-Ozcan et al. (2015a) highlight that "*there is again heterogeneity across countries: European countries for example, are better covered since the mid to late 1990's; many other countries, on the other hand, do not have a significant coverage until 2005-2007. Overall, for most countries, the sample expands over the period 1995-2005, and becomes more or less a stable panel afterwards.*"

Besides financial information, this module also includes more **static descriptive statistics of firms with respect to their identification, contact information, location and sector of activity**. In particular, the descriptive items include, among others, official national identification numbers, address (country, region, city, postcode, street), legal form, year of incorporation, firm status (active, liquidation, merger-acquisition), listed/not listed indicator, industry and activity codes (4 digit level).

[1] For more information about the structure of BvD ID numbers we refer to the following website: https://help.bvdinfo.com/mergedProjects/68_EN/Data/IDNumbers/bvdid.htm.

Table 3 provides an overview of the type of data that is available in the financial module of ORBIS, both in terms of time-series of balanced sheet items and the more static descriptive information needed to identify firms.

Table 3: Financial information in ORBIS

Identification	Accounting classification
- BvD ID number	- Consolidation code
	- Filing type
	- Closing date

Assets	Liabilities and owners' equity
- Tangible fixed assets	- Shareholders funds and capital
- Intangible fixed assets	- Current and non-current liabilities
- Current and total assets	- Long term debt
- Stock	- Loans

Income statement	Financial ratios and margins
- Financial revenues	- Profit margin (%)
- Operating revenue (turnover)	- Gross margin (%)
- Export revenue	- EBITDA margin (%)
- Gross profit	- Current ratio
- Financial expenses	- Liquidity ratio
- R&D expenses	- Solvency ratio
- Taxation	- Operating revenue per employee
- Depreciation & Amortization	- Costs of employees / Operating revenue

Note: The table presents an overview of the type of financial information that is available in ORBIS. This overview is by far not exhaustive but aims to highlight the most relevant financial information for this project.
Calculations: JRC

2.2.2 Ownership information

Data in the **ownership module in ORBIS contains information on a firms' equity ownership information**. Data is collected from BvD's information providers and supplemented by own research from BvD via stock exchanges, newswires and direct contact with companies.

The following data is provided in this module: the name of owners, the BvD ID number to uniquely identify firms, the respective ownership shares, the level of ownership (both direct and total), the type of relation, the country of origin, the source of the information, and the information date. In contrast to the online platform access of ORBIS that only provides ownership information for the latest available year, the historical disks of ORBIS allow to track changes in the ownership structure over time. To this purpose, historical disks contain files with yearly firm ownership structures. Each observation in these files contains information on the link between a target company and its owner or subsidiary.

The type of relation in the ownership module includes simple shareholder, domestic ultimate owner (DUO), and global ultimate owner (GUO). The type of relation constitutes an important variable of the database as it allows filtering searches and collecting for instance all firms with the same parent or identifying all global or domestic ultimate owners of a set of target firms. For an extensive and detailed overview of the

information that is available in the ownership module we refer to Kalemli-Ozcan et al. (2015a).

2.3 Matching procedure

Matching between the participants of the Innovation Radar and ORBIS has been done in several steps. First, the Innovation Radar database has been linked to the COmmon Research DAtawarehouse (CORDA) to obtain more information about the participants of FP projects. CORDA is collecting proposal, evaluation and grant management data of the Framework Programmes, ranging from H2020 to the first Framework Programme FP5. Although most information of CORDA is not freely accessible, it provides the necessary information to link IR participants to ORBIS: VAT identification numbers.

The linkage between the Innovation Radar and CORDA is ensured through the Participant Identification Code (PIC) number, a number of nine digits that uniquely identifies participants in CORDA and the Innovation Radar. **The linkage with CORDA allows to assign a unique VAT number to most participants of the Innovation Radar.** Roughly 85 percent of the participants are associated with a VAT number, while the remaining 15 percent are blank.

The second step is the actual matching between the Innovation Radar participants and ORBIS, based on the VAT number to which each IR participants has been associated. This resulted in the **following two matching procedures**:

- Matching based on VAT numbers;
- Matching based on manual retrieval.

Both methods are described in more detail below. **Before starting any linkage between Innovation Radar and ORBIS, two important caveats should be taken into account**:

- It is important to notice that the PIC number ensures proper linkage back to IR participants and hence it should always be kept as identifier when retrieving information from ORBIS;
- As participants may participate more than once to FP projects, many of them appear multiple times. Hence, to limit the computation time and matching effort, the linkage between Innovation Radar and ORBIS could only be done for unique participants (i.e. by deleting duplicate PICs).

2.3.1 Matching based on VAT numbers

To ensure a proper matching based on VAT numbers between the Innovation Radar and ORBIS, **VAT numbers should be harmonised** in both databases. This problem arise from the fact that VAT numbers may slightly differ in both databases due to differences in the use of lower or upper cases of characters, spaces, and punctuation marks. Hence, all the VAT numbers of the Innovation Radar are harmonised in the following way:

- Transferring all characters to upper case;
- Removing non-alphabetical characters, such as spaces and punctuation marks;

Following these rules, harmonised VAT numbers will only contain alphabetical characters and numbers. This harmonisation is needed to ensure an exact matching between VAT numbers of Innovation Radar and ORBIS.

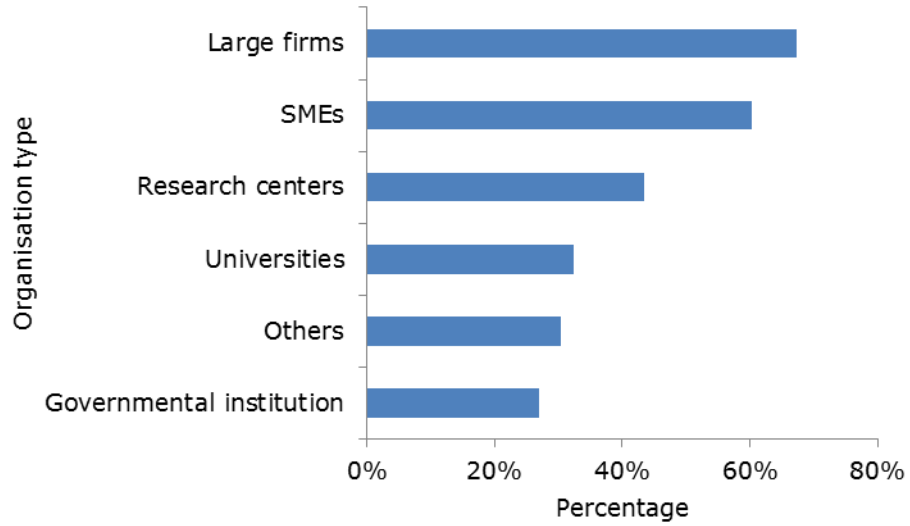
A similar harmonisation method is conducted for ORBIS. As ORBIS contains various firm identifiers for each BvD ID number, the harmonisation procedure is done on each of them, being: national ID number, the national VAT tax identifier, the European VAT number.

As it is a priori not clear to which ORBIS firm identifier the VAT number in the Innovation Radar may refer to, we match the Innovation Radar VAT numbers to each of the ORBIS identifiers mentioned above. Concretely, **we match the harmonised VAT number of the Innovation Radar with the following harmonised identifiers in ORBIS**:

- National ID number;
- National VAT tax identifier;
- European VAT number.

Overall, the matching based on VAT numbers allows to match around 48 percent of the organisations. Figure 1 presents the matching result based on VAT numbers (in percentage) between the Innovation Radar database and ORBIS, by organisation type. Analysing the situation by organisation type, the category of large firms obtain the best matching score with 67 percent of them matched on VAT numbers, while SMEs obtain a 60 percent match. Not surprisingly, remaining organisation types have a lower match with respectively 44 percent for research centers, and around 30 percent for each of the remaining organisation types, being universities, governmental institutions and others.

Figure 1: Matching result based on VAT numbers by organisation type



Note: The figure presents the matching result based on VAT numbers (in percentage) between the Innovation Radar database and ORBIS, by organisation type. Percentages calculated on a total of respectively 1871 SMEs, 1322 large firms, 690 universities, 671 research centers, 412 governmental institutions, and 335 other organisation types.
Calculations: JRC

For the observations that matched with ORBIS, the following variables have been downloaded:

- BvD ID number (key identifier internally used in ORBIS);
- Firm identification numbers (i.e. national ID number, the national VAT tax identifier, the European VAT number);
- International firm name;
- Address information (country, region, city, street).

At this stage, the downloaded database will contain multiple entries for each BvD ID and will require a de-duplication of the database (i.e. obtaining a database with only one entry per BvD ID). In addition, the matched organisations based on VAT numbers

require a validity control to confirm the correctness and accuracy of the matching procedure. Both the processes of de-duplication of the database and of validity control are described below.

De-duplication of the database

The main reason for the appearance of multiple observations per BvD ID numbers have been highlighted by Kalemli-Ozcan et al. (2015a). It stems from the fact that countries use more than one type of national identifier. Hence, observations will appear multiple times according to the number of national identifiers that are used. This will lead to same BvD IDs to be linked to different national identifiers. The opposite case of similar national identifiers linked to multiple BvD IDs may occur as well. This can be the case when a company contains multiple branches or changed address information.

As the de-duplication is not always straightforward to automatise, the removal of duplicates has been done manually on a case-by case basis. The percentage of IR organisations that matched with ORBIS and had multiple observations in ORBIS remained relatively limited. De-duplication was needed for roughly eight percent of the sample of matched organisations.

Matching validity control

After the de-duplication of the database, a validity control has been conducted to assert that the matched observations have been matched with the correct organisation. To this purpose, address information and organisation names has been downloaded from ORBIS in order to compare them across the Innovation Radar database and ORBIS.

For the validity control, similarity measures have been created for the organisation name, postcode, street, city and country.[2] Comparisons of these similarity measures across the Innovation Radar database and ORBIS allowed to assess the matching result. In most of the cases, consistency was found, which validated the matching result. For roughly 50 cases, the BvD ID was changed or reverted to zero.

2.3.2 Matching based on manual retrieval

Manual matching was needed to solve any matching problem encountered during the matching with VAT numbers and to retrieve participants' information in ORBIS for which no VAT numbers were found in CORDA. As manual matching is time consuming, this matching procedure was limited to the firms that were not matched yet. Hence, roughly 1300 firms (i.e. SMEs and large firms) have been matched manually. This matching procedure was relatively successful as 88 percent of the unmatched firms could be retrieved manually in ORBIS.

2.4 Data coverage

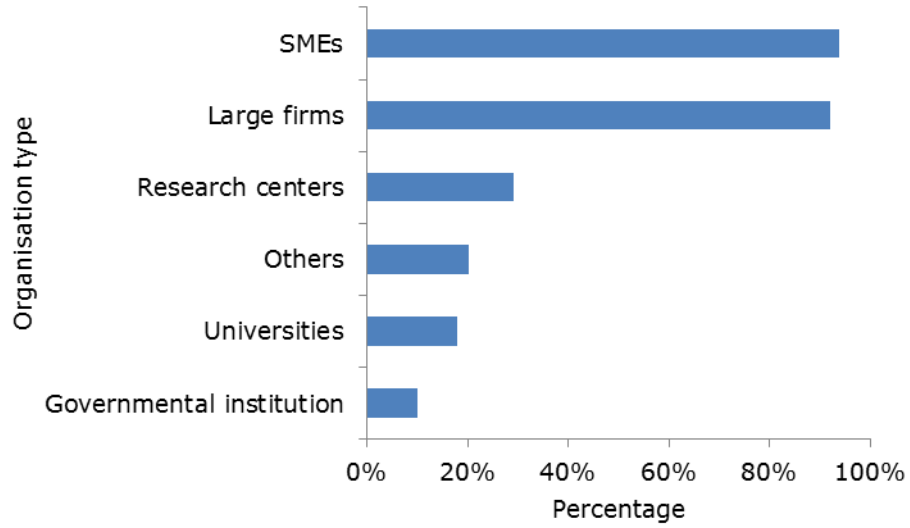
Following the matching procedure outlined above, IR organisations could be matched with their corresponding BvD ID in ORBIS which allows to uniquely identify these organisations in ORBIS and to download their financial information.

Figure 2 presents the final matching result between the Innovation Radar database and ORBIS and highlights by organisation type the percentage for which financial information was available. Analysing the situation by organisation type, the category of SMEs obtain the best matching score with 94 percent of them matched, while large firms obtain a 92

[2] The creation of similarity measures was based on the Matchit routine from STATA, holding all options on their default.

percent match. Provided that ORBIS is mainly covering firms, no manual matching has been performed on the remaining organisation types. Hence, their matching with financial information of ORBIS is significantly lower. The remaining organisation types have a matching percentage that varies between 10 and 30 percent.

Figure 2: Final matching result with ORBIS by organisation type



Note: The figure presents the final matching result (in percentage) between the Innovation Radar database and ORBIS, by organisation type. Percentages calculated on a total of respectively 1871 SMEs, 1322 large firms, 690 universities, 671 research centers, 412 governmental institutions, and 335 other organisation types.
Calculations: JRC

The BvD ID linked to the IR organisations can then be used to download static and dynamic time-series data from the financial module of ORBIS.

2.4.1 Descriptive information

Table 4 provides an overview of the descriptive information that has been downloaded from ORBIS for the IR organisations that matched with ORBIS. While most information in this table is straightforward, we comment some of them that might not be directly obvious. Variables related to the industry classification provide information about the economic activities of firms. NACE is the industry standard classification system used in the European Union. The current version at use is revision 2.[3] Firms can report multiple industries. However, as reported in the table below, only primary/core NACE codes (i.e. the main industry of a firm) has been downloaded. The firm status refers to the current level of activity a firm, being active, in liquidation or acquired/merged, among others (see Section 2.5.2 for more details on this variable). The entity type classifies organisations according to the nature of their activity, distinguishing – among others – between industrial companies, banks, financial or insurance companies, public authorities and governments, and research institutes. Firm category is a firm size classification developed by BvD and distinguishing between small, medium sized, large and very large firms. Finally, the IPO date refers to the Initial Public Offering (IPO) date on which a firm offers its stock to the public for the first time.

[3] For a detailed list of the NACE Rev. 2 classification we refer to the following website of Eurostat: http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=.

Table 4: Descriptive information downloaded from ORBIS

Identification	Industry classification
- BvD ID number	- NACE Rev. 2, Core code (4 digits)
- National ID number	- NACE Rev. 2, Primary code(s)
- European VAT number	- NACE rev.2, primary code , text description

Contact information	Legal information
- International firm name	- Firm status
- Address:	- Date of incorporation
- Country ISO code	- Type of entity
- NUTS 2 and NUTS 3 region	- Firm category
- Postcode and city	- IPO date
- Street	
- Telephone number	
- Website	
- E-mail address	

Note: The table provides an overview of the descriptive information that has been downloaded from ORBIS.
Calculations: JRC

2.4.2 Financial information

Prior to downloading any financial information, it is important to note that ORBIS uses a range of variables to link the various balance sheet items to a firm. The range of variables that identify each balance sheet item are: BvD ID number, the consolidation code, the filing type and the closing date. Hence, the combination of all these variables together identifies each balance sheet item in the ORBIS database. To clarify this identification key for the balance sheet items, each of these variables are discussed in more detail below.

BvD ID number

As explained above, this number is needed to uniquely identify an organisation in ORBIS and hence it should be retrieved at each download as it constitutes the key identifier in ORBIS and all other databases owned by BvD.

Consolidation code

The consolidation code in ORBIS identifies the type of financial account that is reported. ORBIS distinguishes between various consolidation types, being:

- Consolidated account C1: account of a company-headquarter of a group, aggregating all companies belonging to the group (affiliates, subsidiaries, etc.), where the company-headquarter has no unconsolidated account;
- Consolidated account C2: account of a company-headquarter of a group, aggregating all companies belonging to the group (affiliates, subsidiaries, etc.) where the company-headquarter also presents an unconsolidated account;
- Unconsolidated account U1: account of a company with no consolidated account;
- Unconsolidated account U2: account of a company with a consolidated account;
- Limited number of financial items LF: account of a company with only a limited number of information/variables included;

- No financial items at all NF: account of a company with no financial items/variables included;

In general, ORBIS makes a distinction between consolidated accounts (the statement of a parent firm integrating the statements of its controlled subsidiaries) and unconsolidated accounts (the statement not integrating the statement of the controlled entities). Some (large) firms report only unconsolidated accounts, or only consolidated ones or a combination of both. In addition to these types of accounts, ORBIS contains account items labelled as limited or no recent financials. In most of these cases, only the operating revenue and number of employees are available. Hence, a same firm may report account items that belong to different consolidation categories. As stated by Kalemli-Ozcan (2015b, p. 20), *"...The type of account reported is related to country filing requirements for particular size or the legal type of companies, as detailed in Table A.1*".

Filing type

ORBIS can obtain financial statement information through different channels, depending on the type of filing firms are using to report their financial results. Two filing types are distinguished in ORBIS:

- Annual report;
- Local registry.

When downloading data from the financial module, a firm may contain account items for both filing types. However, due to different reporting rules, same financial variables may have different values across the two filing types (Kalemli-Ozcan et al., 2015a).

Closing date

The closing date defines the cycle period of the accounting exercise and is expressed in day-month-year format in ORBIS. In order to create an accounting year variable, we substract the year information of the closing date. A limited number of firms change their closing date over time (e.g. change from May to December). In this case, reports may be given for both months during the year of change.

From the above explanation, it is obvious that the combination of BvD ID number, consolidation code, filing type and closing year is needed to identify to which type of accounting statement a financial variable is referring to. However, all these differences in accounting reports bring along problems of duplication. Below we describe the convention rules we used to de-duplicate the accounting information.

De-duplication of accounting information

For the de-duplication of the accounting information, we prioritise certain accounting information while deleting others, along the following rules:

- In case a firm contains accounting items with similar consolidation codes but information for different filing types, we give priority to the Local registry filing;
- In case a firm contains both unconsolidated and consolidated accounting items, we give priority to the unconsolidated accounting information;
- In case a firm contains both consolidated account information and accounting items labelled as limited financial information, we give priority to the consolidated accounting information;

- In case a firm contains duplicates in BvD ID-Year variable combination due to changes of the closing date (see above), we randomly delete duplicate observations;
- In case a firm contains multiple filing types and one of them contains more information (e.g. number of employees is missing in the local register but available in the annual report), this information has been duplicated to the one that has been kept in the final database.

To summarise, Table 5 presents an overview of the financial information that has been downloaded from ORBIS. At the moment of writing this report, the selection of financial information remains relatively limited but can be easily extended to any other balance sheet item presented in Table 3.

Table 5: Financial information downloaded from ORBIS

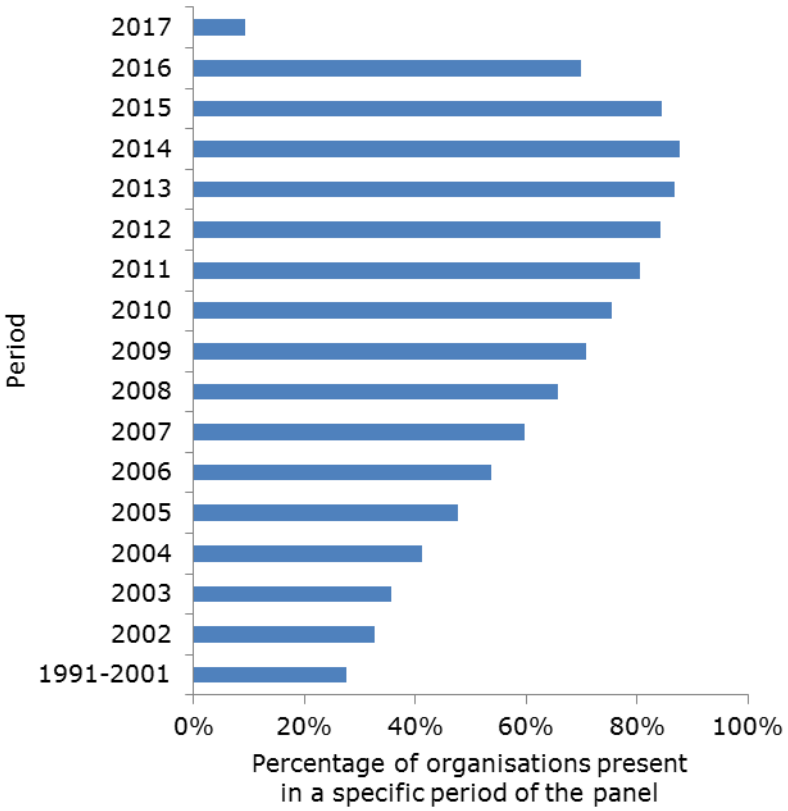
Identification	Accounting classification
- BvD ID number	- Consolidation code
	- Filing type
	- Closing date

Key financials	Trade and R&D
- Number of employees	- Export revenue
- Operating revenue (Turnover)	- Research & Development expenses
- Enterprise value	

Note: The table provides an overview of the financial information that has been downloaded from ORBIS.
Calculations: JRC

To obtain a better understanding of the time representation of the IR organisations that have been matched with ORBIS, Figure 3 displays the percentage of organisations that are present in the panel, by period. This figure does not provide an overview of the data coverage for the different balanced sheet items that have been downloaded (i.e. specific balance sheet items may still be missing for many observations), but it provides an overview of the years covered in the panel. The time representation is in line with the two-year time lag of ORBIS as mentioned by Kalemli-Ozcan (2015b). Only a limited percentage of organisations has information for 2017, while 2018 is not covered in the database. The largest coverage is observed in the period 2010-2015 with percentages around 70-80, while the coverage gradually decreases over time. The gradual increase over time is due to a lower coverage in ORBIS for earlier years and to the decreasing frequency of older firms in the sample.

Figure 3: Years covered in the panel of financial information downloaded from ORBIS



Note: The figure presents the percentage of organisations that are present in the panel, by period. This figure does not provide an overview of the data coverage for the different balanced sheet items that have been downloaded (i.e. specific balance sheet items may still be missing for many observations), but it provides an overview of the years covered in the panel.

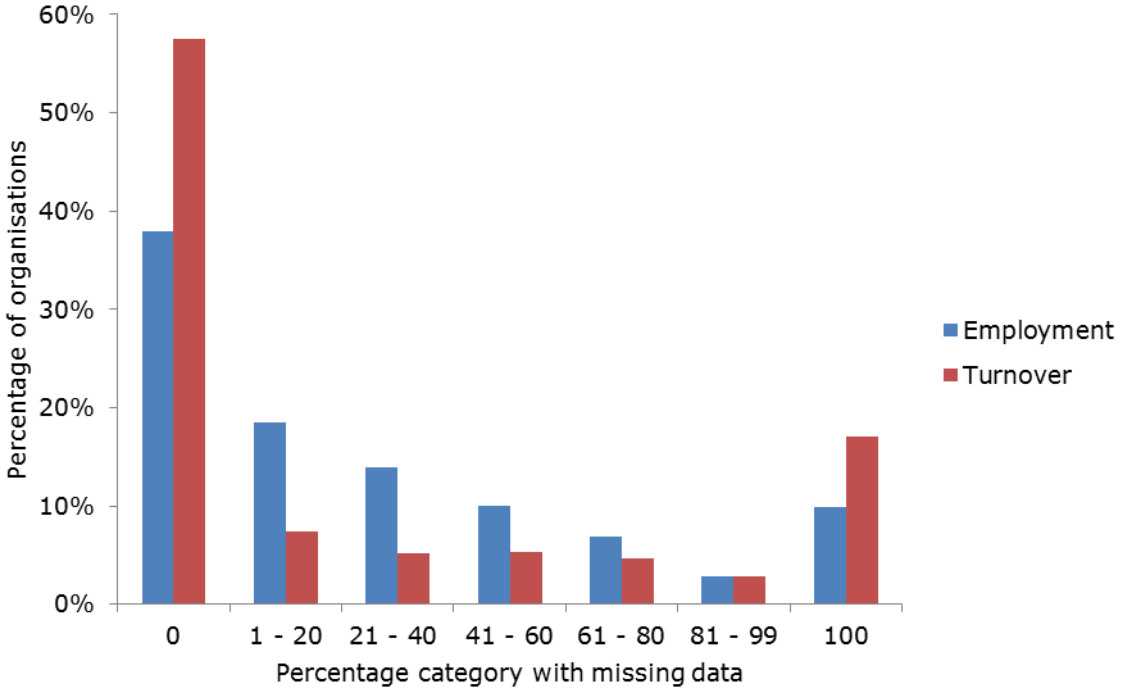
Calculations: JRC

To analyse missing data patterns in the panel structure for employment and turnover, we calculate for both balance sheet items the percentage of values for each organisation that are missing since the appearance of the organisation in the downloaded database. In case an organisation appears for 10 years and has 4 missing values on employment while other observations are filled in, it will get allocated a missing employment score of 40 percent. Similar calculations are conducted for turnover.

Figure 4 displays the percentage of organisations with missing information on employment and turnover, by different levels of missing patterns. We distinguish between different percentage groups of missing data: 0 (i.e. no missing data in the panel), 1-20 percent of missing data, and subsequent percentage groups up to 100 (i.e. all years have missing data). Figure 4 reveal the following:

- Roughly 40 to 60 percent of the matched organisations have data on respectively employment and turnover across all years;
- Roughly 10 to 20 percent of the matched organisations have missing data on respectively employment and turnover across all years;
- In general, the percentage of organisations with missing data, gradually decreases for higher levels of missing data.

Figure 4: Percentage of missing data for employment and turnover



Note: The figure presents the missing patterns of employment and turnover in the panel data. It displays the percentage of organisations with missing information on employment and turnover, by different levels of missing patterns. We distinguish between different percentage groups of missing data: 0 (i.e. no missing data in the panel), 1-20 percent of missing data, and subsequent percentage groups up to 100 (i.e. all years have missing data).

Calculations: JRC

2.5 Data analysis

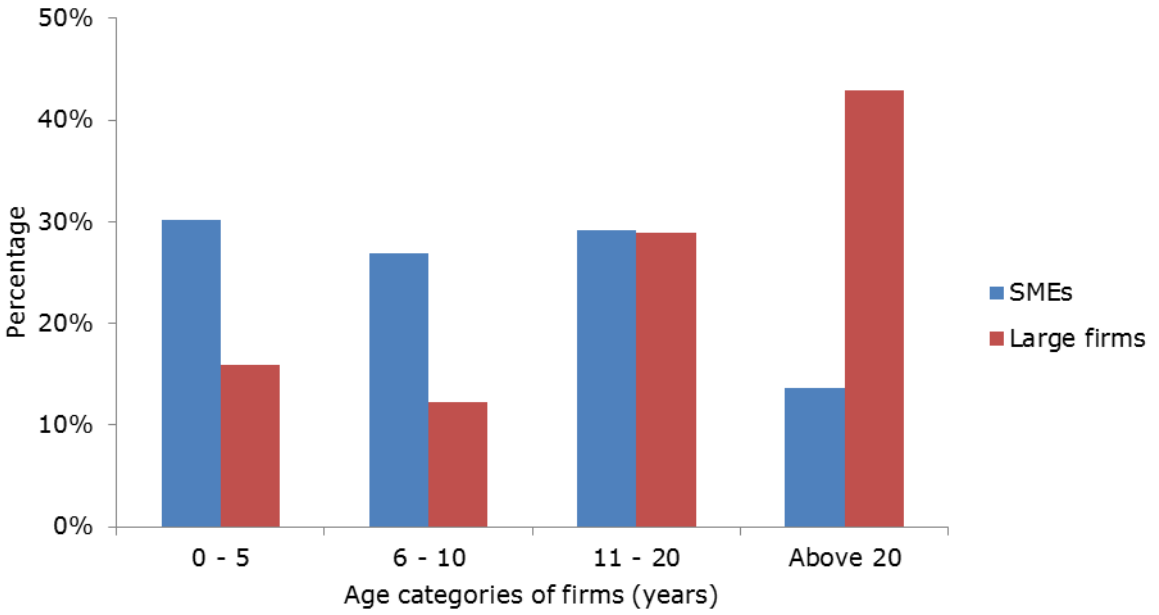
In this section, we provide some preliminary data analyses based on the matched database between Innovation Radar and ORBIS. The data analysis in this section is restricted to firms. Hence, we only provide statistics for SMEs and large firms. The main reason for this restrictive analysis is that ORBIS is primarily screening firms and has more limited data about other organisation types such as universities, research centers and governmental institutions.

2.5.1 Age of firms

Figure 5 presents the percentage of firms by age category. The age of a firm is calculated at the first participation occurrence of a firm to a FP project scanned by the Innovation Radar. Around 30 percent of SMEs are between zero and five years old. A similar percentage score is observed for age categories of 6-10 and 11-20 years old, revealing that a large part of SMEs are relatively old. The percentage of SMEs older than 20 years drops to 14 percent.

A different pattern is observed for large firms, where the majority is situated in the age category 11-20 years old or older than 20 years. Around 12 to 15 percent of large firms are relatively young, being in the age categories of 0-5 and 6-10.

Figure 5: Age categories of firms

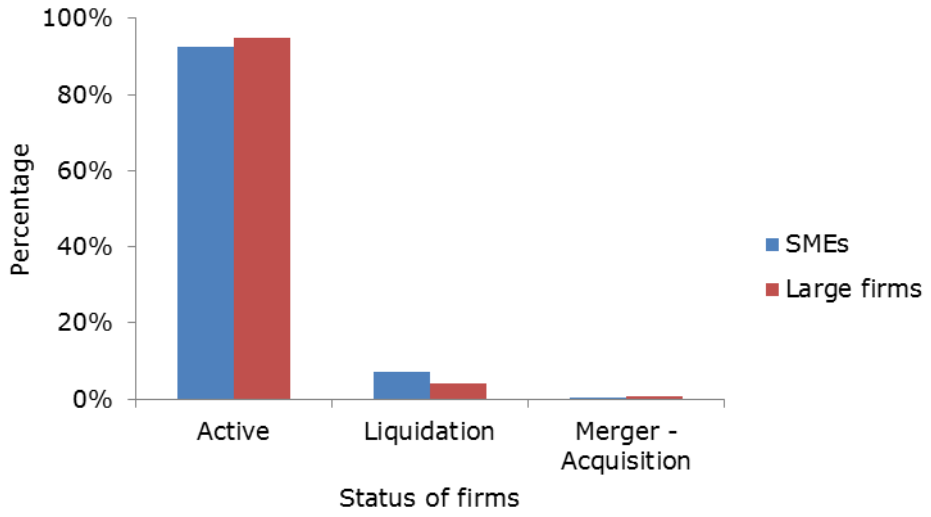


Note: The figure presents the percentage of firms by age category. The age of a firm is calculated at the first participation occurrence of a firm to a FP project scanned by the Innovation Radar.
Calculations: JRC

2.5.2 Status of firms

In order to control the actual status of firms, ORBIS keeps the latest available information about a firm's activity, distinguishing among others between firms that are active, in liquidation and merged/acquired. Figure 6 presents the status of both SMEs and large firms. The figure reveals that the large majority of firms (above 90 percent) is still active. Only four percent of firms are in liquidation and less than one percent of firms has been merged or acquired.

Figure 6: Status of firms



Note: Percentages based on respectively 1767 SMEs and 1242 large firms. In ORBIS, the variable Status takes the values Active, Active (default of payments), Active (dormant), Active (insolvency proceedings), Active (rescue plan), Bankruptcy, Dissolved, Dissolved (merger or take-over), Dissolved (liquidation), Dissolved (bankruptcy), In liquidation, Inactive, Inactive (no precision), Unknown. The firm status categories in this report are classified as: Active (i.e. Active status), Merger-Acquisition (i.e. Dissolved (merge or take-over)) and Liquidation (all other status values).

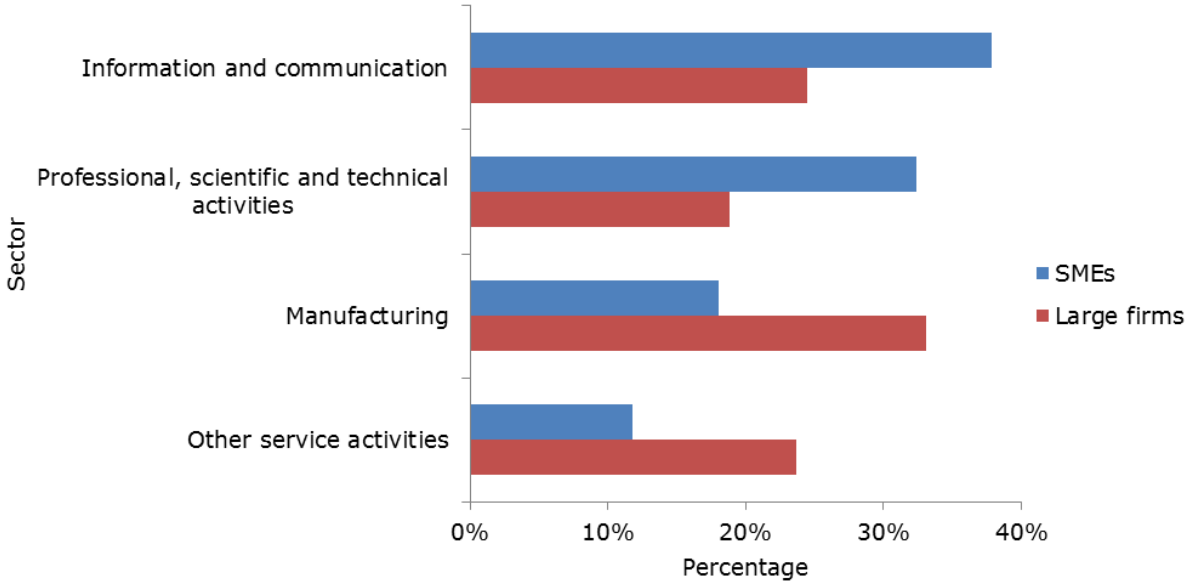
Calculations: JRC

2.5.3 Sector classification

Figure 7 presents the sector classification for SMEs and large firms. As the current version of the Innovation Radar covers only FP projects with an ICT theme, it is not surprising to note that nearly 40 percent of SMEs are active in the information and communication sector. Roughly 30 percent of SMEs are operating professional, scientific and technical activities, while percentages of SMEs in manufacturing and other service activities drop down to 20 and 10 percent.

Large firms show a different pattern, with the highest percentages reported in manufacturing (33 percent). While sectors of information and communication, and other service activities have a similar distribution (around 24 percent), the lowest percentage of large firms is reported in professional, scientific and technical activities

Figure 7: Sector classification



Note: Percentages based on respectively 1760 SMEs and 1240 large firms.
Calculations: JRC

2.5.4 Employment

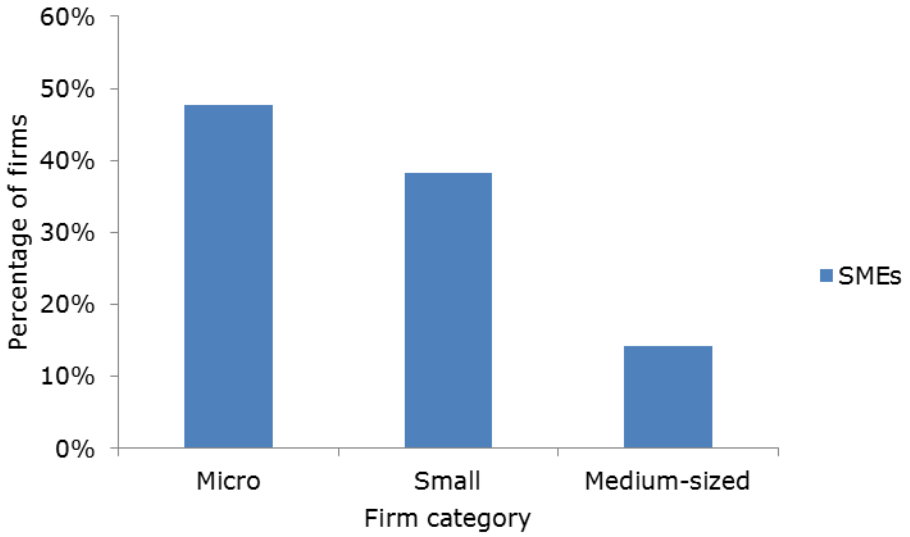
In terms of employment, both the employment level and growth are analysed to obtain a better understanding of the company size and growth across different firm categories. In this paragraph we mainly focus on statistics for SMEs for two reasons. First, SMEs have a higher degree of employment dynamics than large firms and represent the most interesting subsample as these firms are often perceived as the driving forces behind innovations. Second, the statistics of large firms may provide a biased representation as employment information has been retrieved for unconsolidated accounts. Hence, employment statistics of large firms represent employment figures of the headquarter and not the integrated statements of its controlled subsidiaries.

Employment level

To explore company sizes in terms of employment level, we classify SMEs in three categories, according to the following firm categories as defined by the European Commission (EC, 2018):

- Micro firms have up to 10 employees;
- Small firms have up to 50 employees;
- Medium-sized firms have up to 250 employees.

Figure 8: SMEs distribution across employment levels



Note: The figure presents the distribution of micro, small and medium-sized firms (as identified by their employment level) in the sample of SMEs screened by the Innovation Radar.
Calculations: JRC

Figure 8 presents the distribution of micro, small and medium-sized firms in the sample of SMEs screened by the Innovation Radar and as identified by their employment level. This figure reveals the following:

- Half of the sample of SMEs are micro-firms in terms of employment;
- Around 14% of SMEs are medium-sized firms in terms of employment.

Employment growth

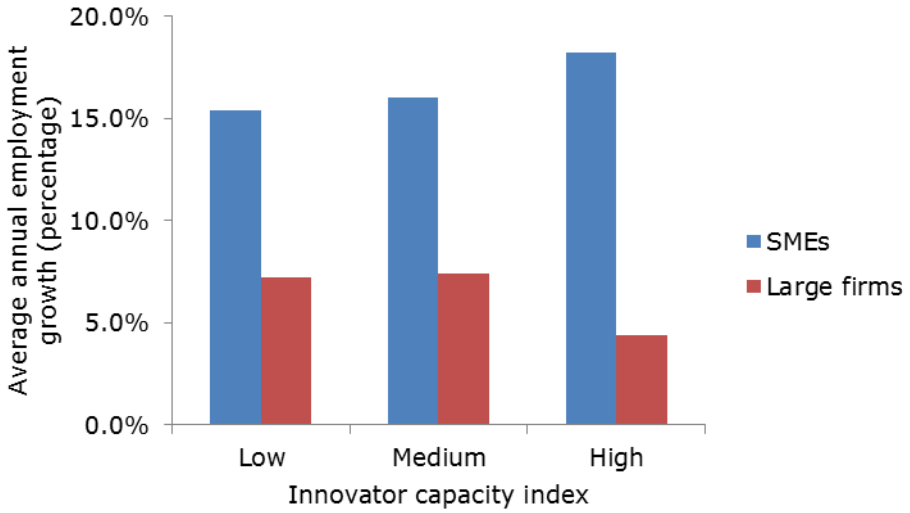
Figure 9 presents the annual employment growth for different levels of innovator capacity. The innovator capacity of organisation participating to FP projects screened by the Innovation Radar is measured with the Innovator Capacity Index as highlighted on page 14 in a report of De Prato et al. (2015). In line with that report (see page 14) we identify three different levels of innovator capacity: low, medium and high.[4]

Figure 9 reveals the following:

- Annual employment growth gradually increases for higher levels of innovator capacity;
- The annual employment growth of innovators with a high innovator capacity is 3 percentage points higher than that of other organisations.

[4] Organisations that have not been identified by the Innovation Radar as key innovators have been assigned an innovator capacity index score of zero.

Figure 9: Annual employment growth by levels of innovator capacity



Note: The figure presents breakdowns of annual employment growth averages by firm category and for different levels of the Innovator capacity index. Only the highest innovator capacity index score is retained for innovators having multiple ones. To avoid biases in employment growth figures due to outliers, most extreme cases (i.e. growth increases above 500 percent or decreases above 100 percent) are not taken into account.
Calculations: JRC

2.5.5 Turnover

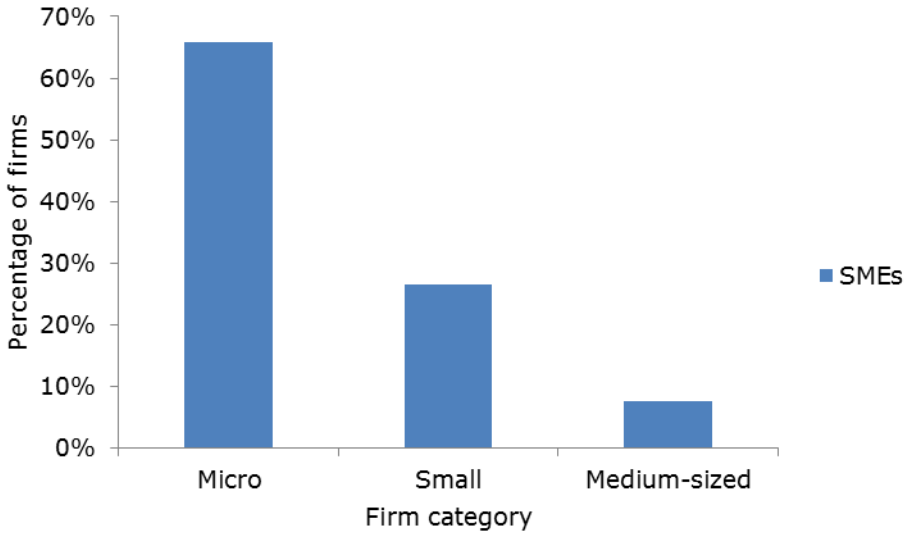
In a similar vein as for employment, turnover is analysed both at level and growth rates. This paragraph presents the main trends in terms of turnover for the SME sample of FP projects screened by the Innovation Radar.

Turnover level

To explore company sizes in terms of turnover level, we classify SMEs in three categories, according to the following firm categories as defined by the European Commission (EC, 2018):

- Micro firms have up to 2 million euros;
- Small firms have up to 10 million euros;
- Medium-sized firms have up to 50 million euros.

Figure 10: SMEs distribution across turnover levels



Note: The figure presents the distribution of micro, small and medium-sized firms (as identified by their turnover level) in the sample of SMEs screened by the Innovation Radar.
Calculations: JRC

Figure 10 presents the distribution of micro, small and medium-sized firms in the sample of SMEs screened by the Innovation Radar and as identified by their turnover level. This figure reveals the following:

- Roughly 65% of the sample of SMEs are micro-firms in terms of turnover;
- Around 8% of SMEs are medium-sized firms in terms of turnover.

Turnover growth

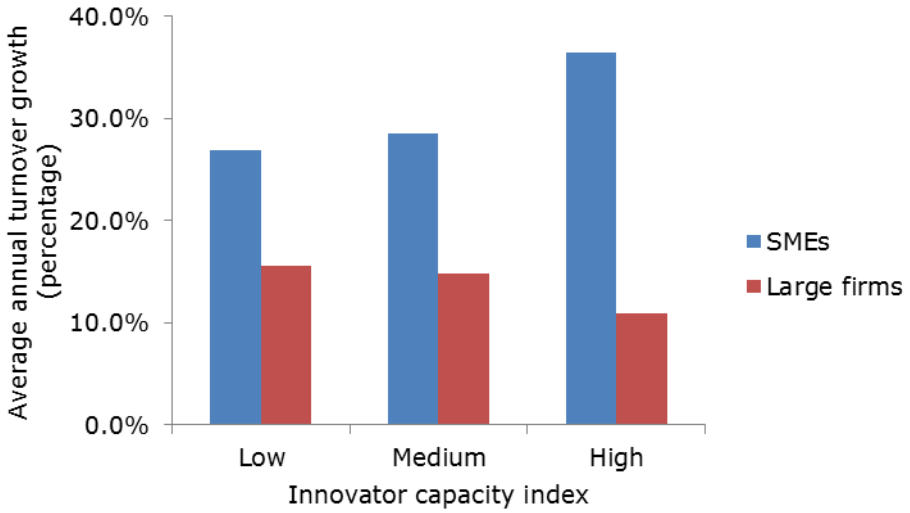
Figure 9 presents the annual turnover growth for different levels of innovator capacity. The innovator capacity of organisation participating to FP projects screened by the Innovation Radar is measured with the Innovator Capacity Index as highlighted on page 14 in a report of De Prato et al. (2015). In line with that report (see page 14) we identify three different levels of innovator capacity: low, medium and high.[5]

Figure 9 reveals the following:

- Annual turnover growth gradually increases for higher levels of innovator capacity;
- The annual turnover growth of innovators with a high innovator capacity is 8 to 10 percentage points higher than that of other organisations.

[5] Organisations that have not been identified by the Innovation Radar as key innovators have been assigned an innovator capacity index score of zero.

Figure 11: Annual turnover growth by levels of innovator capacity



Note: The figure presents breakdowns of annual employment growth averages by firm category and for different levels of the Innovator capacity index. Only the highest innovator capacity index score is retained for innovators having multiple ones. To avoid biases in turnover growth figures due to outliers, most extreme cases (i.e. growth increases above 800 percent or decreases above 100 percent) are not taken into account.
Calculations: JRC

2.6 Alternative firm-level databases

A wide range of alternative firm-level databases exist containing financial information of firms. We provide a non-exhaustive list of alternative databases:

- Worldscope (Thomson Reuters);
- Dun & Bradstreet;
- LexisNexis;
- Open source company databases based on web scraping.

However, ORBIS is among the most widely used firm-level databases with improving coverage over time. Many alternative databases have biased coverage (towards large firms), and are not as exhaustive as ORBIS.

3 Patent information from PATSTAT

In this chapter, we describe the process and results of the linkage between the Innovation Radar and the EPO PATSTAT Worldwide Patent Statistical Database by matching Innovation Radar beneficiary names with PATSTAT patentee names.

First we describe the PATSTAT database, next we discuss the method we developed to match organisation names within Innovation Radar with PATSTAT patentee names, and finally we elaborate on the matching results.

3.1 Access to PATSTAT

The PATSTAT database is made available online and offline as a set of text files that can be loaded into a relational database of statistical program (one file for every table within the database: application information, publication information, inventor and applicant information, classification information, …).

3.2 PATSTAT data

The PATSTAT database offers worldwide coverage of patent applications and procedural data (application data, publication data, grant data) and legal status records.

It contains more than 100 million patent records of about 90 patent issuing authorities – including major offices as USPTO, EPO, WIPO, JPO - from mid-19th century up to today and is updated twice a year (sprint and autumn edition).

PATSTAT version 2018 spring is used for this matching project. The PATSTAT database is not freely accessible but requires an annual subscription. For more information about subscription options we refer to the PATSTAT website.

3.3 Matching procedure

Organisation name matching is not straightforward due to the many spelling variations and errors in organisation names as registered in databases. As such, approximate (or "fuzzy") string matching techniques are necessary to cope with these variations and find similar names. Many approximate string searching techniques are available (e.g. Levenshtein distance, Jaccard similarity, Jaro-Wikler distance) but their application on large datasets is challenging because every potential combination of a source and target name has to be evaluated, resulting in an exponential growing calculation time.

To deal with this computational challenge, a two-step approach is used. First, for every Innovation Radar name a list of related PATSTAT patentee names is compiled using a very fast – but inaccurate - approximate string searching technique. This is the search step. In this step a method is used that is able to quickly identify related names based on an overestimating similarity measure. I.e. almost all items that are somewhat related will be found, but the majority of found items will be not related (because of the use of a similarity measure that is very fast to calculate be largely overestimates the real similarity). In technical terms, this method is a high recall / low precision method: almost all related names will be found, but most of the related names will turn out to be unrelated in practice).

Next, as most of the potential matches of step one have an insignificant relatedness, an assessment has to be made whether the proposed related names are indeed pointing to the same underlying company or organisation. This is the validation step. In this step, address information is used to assess the validity of the proposed related names from step 1.

3.3.1 Step 1: Matching based on three-level significance of terms

In this step a list of related names is compiled. As the PATSTAT dataset contains millions of patentee names, a fast approach is needed that identifies as many related names as possible. A three-level approach is used to identify as much items as possible in an efficient way. First, all source names from the Innovation Radar dataset and all target names from PATSTAT are split into terms and a list is compiled with all terms appearing in the source and target names. Next, the target document frequency is derived for every term, i.e. the number of target names (PATSTAT patentee names) that contain that particular term. Based on this document frequency, all terms are classified as having a high significance (class 1, terms with low document frequency), a medium significance (class 2, terms with medium document frequency), a low significance (class 3, terms with high document frequency), or no significance (class 0, terms with very high document frequency).

The idea is the following: the less target names contain a particular term, the higher the probability that two names sharing that one term are indeed related (e.g. term "STMICROELECTRONICS" does not occur in a lot of PATSTAT names, hence all Innovation Radar names containing the term "STMICROELECTRONICS" and all PATSTAT names containing the term "STMICROELECTRONICS" have a high probability to be related. On the other hand, the more target names contain a particular term, the lower the probability that two names sharing that one term are indeed related (e.g. term "international" does occur in a lot of PATSTAT names, hence the fact that an Innovation Radar name and a PATSTAT name both contain the term "international" is not enough to decide that these names are related).

Based on this idea, a list of potential matches is derived at three levels as following: a first subset is compiled using all source and target names sharing 1 high significant term (i.e. term with low document frequency); a second subset is compiled using all source and target names sharing 2 medium significant terms (i.e. terms with medium document frequency); and a third subset is compiled using all source and target names sharing 3 medium or low significant terms (i.e. terms with medium or high document frequency respectively).

This allows a very fast compilation of lists with almost all related names.

3.3.2 Step 2: Validation based on name and address similarity

The nature of the matching step (identify related names based on one, two or three matching terms) results in many potential matches that are unrelated in practice (many names sharing one or two or three terms are still unrelated).

The potential matches identified in step 1 are assessed using name and address similarity. First, more fine-grained names similarity matches are used. For all potential matches (every pair of source name and potential related target name from step 1), Jaccard, Dice and overlap similarity measures are calculated based on the bigrams present in the source and target name (we deliberate do not use edit-distance based methods as Levenshtein and Jaro-Winkler as these depend on the character order and would result in too many false negatives for organisation name matching). These techniques result in a more reliable name similarity assessment, but are too slow to run on all name combinations (hence we first use a far faster but less accurate method in the first step). However, even with those more fine-grained measures, it is still impossible to clearly distinguish related names pointing to identical organisations from related but non-identical organisation; although we can set a clear minimum and maximum threshold for matching names: source and target names with Jaccard, Dice or overlap similarity below 50% are as good as never related, and above 90% are almost always identical. But this leaves a large range of cases that cannot be judged on name similarity alone. Hence an additional validation mechanism is needed and the only

additional information available to use in the validation is the address information available in the Innovation Radar and PATSTAT dataset. This approach is however hampered by two issues: first, not all PATSTAT patentees have address data available, and second, the presence of significant address similarity clearly indicates a high probability of matching organisations, but the absence of address similarity does not necessarily mean that there is no match (organisations can have multiple sites).

The procedure to deal with these issues is the following: starting point is a list of potential matches based on the matching in step 1, with a minimum Jaccard, Dice or overlap similarity based on the bigrams in the names. For every potential match, all available addresses of the source and target name are compared side by side. If at least one address of the source organisation has street and city similarity with one address of the target organisation (based on overlap similarity on all bigrams present in the address strings), than the source name is linked with all PATSTAT patentee records with the related target name, under the condition that the country code of the PATSTAT record is identical to that of the source address (similar names with different country cannot be identical in a legal perspective, so records with the similar name but different country code are not linked). Mind that only the country code has to match, not the complete address (but the starting point is that at least one of the addresses has a street and city match). As such the procedure deals with multi-site organisations (if the same patentee name is available in the PATSTAT dataset with multiple addresses, all records with the same country code are linked if at least one street and city matches with the source address).

To deal with patentee records without address, all other addresses are taken into account to decide. If at least one street and city matches with the source address, and all addresses with country codes have the same country code, than the patentee records with the same similar name but no address information are also linked to the source name (the idea is that if they were not related, different addresses for the same name would pop-up). If not all addresses have the same country code, patentee records without address information cannot be allocated, because there is no way to find out to which country they might belong. Essentially, this procedure is equivalent to address replenishment as regularly used with PATSTAT data (fill missing addresses with addresses of similar patentees within patent families).

For potential matches with very high name similarity (Jaccard similarity based on bigrams above 90%, or above 99% for short strings), street and city addresses are not taken into account and these source and target names are directly linked together if they have the same country code (again similar names with different country cannot be identical in a legal perspective). Again for patentees without available address (country) information, all other address are taken into account: potential matches with very high name similarity are linked to patentee records without address country information if there are no other patentee records available with address data, or no other patentee address records available with address countries different from the source address.

3.4 Data coverage

Following the matching procedure outlined above, IR organisations could be matched with their corresponding patent information in PATSTAT. Matching results for the distinct organisations can be summarised as follows:

- For 2214 distinct organisations a match was found in the PATSTAT database (42 %);
- For 3087 distinct organisations no match was found in the PATSTAT database (58%).

Table 6 presents the final matching result between the Innovation Radar database and PATSTAT and highlights by organisation type the percentage of matched organisations

and the number of patent applications that could be retrieved for matched ones. Analysing the situation by organisation type, the category of universities obtain the best matching score with 84 percent of them matched, while research centers obtain a 55 percent match. The matching with patent information is slightly lower for large firms and SMEs, with respective percentages of 47% and 30%. The remaining organisation types, being governmental institutions and other types have a matching percentage that varies between 16 and 6 percent.

Table 6: Final matching result with PATSTAT by organisation type

Organisation type	Total	Matched		Patent applications
Universities	690	580	84.1%	261,459
Research Centers	671	368	54.8%	289,560
Large firms	1322	624	47.2%	1,339,983
SMEs	1871	558	29.8%	13,913
Governmental institutions	412	65	15.8%	4,304
Others	335	19	5.7%	4,335
Total	**5,301**	**2,214**	**41.8%**	**1,913,554**

Note: The table presents an overview of the final matching result of organisations in the Innovation Radar database with PATSTAT, by organisation type.
Calculations: JRC

3.5 Data analysis

In this section, we provide some preliminary data analyses based on the matched database between Innovation Radar and PATSTAT. In particular, we provide some basic statistic descriptives in terms of patent applications and the top 5 technology fields covered by organisations of the Innovation Radar.

3.5.1 Number of patent applications

For the partner organisations having patents, the number of patent applications range from 1 to 172,079 with a mean of 908, a median of 41 (1st quantile = 6; 3rd quantile = 241) and a standard deviation of 6,612.

Table 7 contains summary statistics of number of patent applications by organisation type. In particular it provides the breakdown by organisation type of the minimum, maximum, mean, median and standard deviation of the number of patent applications per partner organisation.

Table 7: Summary statistics of number of patent applications by organisation type

Organisation type	Min	Max	Mean	Median	Q1	Q3	Sd
Universities	1	30,884	487	160	41	431	1,664
Research Centers	1	43,717	849	44	11	234	4,048
Large firms	1	172,079	2,224	80	10	583	11,832
SMEs	1	751	26	7	3	23	60
Governmental institutions	1	1,702	69	3	1	21	234
Others	1	3,812	228	10	1	24	869

Note: The table presents summary statistics of number of patent applications by organisation type. In this table following abbreviations have been used: Q1: First quartile (i.e. middle value between the minimum and the median), Q3: Third quartile (i.e. middle value between the median and the maximum), Sd: Standard deviation.

Calculations: JRC

3.5.2 Technology fields

Table 7 contains summary statistics of the top 5 technology fields covered by the patents of the Innovation Radar organisations that could be matched with PATSTAT. In line with the ICT thematic field of the FP projects that have been screened by the pilot edition of the Innovation Radar, the top 5 includes technology fields linked to semiconductor devices, analysis of materials in determining their chemical and physical characteristics, speech analysis and recognition and electric digital data processing. The only technology field that does not seem to directly relate to ICT is the preparation for medical and dental purposes. This technology field may reflect the patenting activities of some large firms active in that field, and hence are not representative for the wider sample of organisations in the IR.

Table 8: Summary statistics of top 5 technology fields

Rank	Technology field	Patent applications
1	H01L - Semiconductor devices	416,592
2	G01N - Analysing materials determining their chemical and physcial properties	351,390
3	G10L - Speech analysis and recognition	306,977
4	A61K - Preparations for medical and detal purposes	286,240
5	G06F - Eletric digital data processing	225,966

Note: The table presents summary statistics of top 5 technology fields covered by the patents of the Innovation Radar organisations that could be matched with PATSTAT.

Calculations: JRC

3.6 Alternative patent databases

There are other sources of patent information. A non-exhaustive list of alternative databases includes:

- National Patent Offices;
- Google Patent;
- PATENTSCOPE by WIPO.

However, the European Patent Office (EPO), i.e. the publisher of PATSTAT, collects and harmonises raw information on patent applications filed to around 90 Patent Offices, including the EPO itself, the US Patent and Trademark Office (USPTO), the Japan Patent Office (JPO) as well as the other most active Patent Offices worldwide, including China and India. As a result, PATSTAT covers about 99% of the total number of patent applications submitted worldwide. The relatively low price and access to raw data, allows to match patent information with inventor and applicant level data and to compute custom indicators.

4 Venture Capital funding information from Dealroom

In order to provide information of Venture Capital (VC) funding to innovators identified by the Innovation Radar, Dealroom was used as a source of information. It provides comprehensive data on venture-backed and private equity-backed companies – including their investors and executives – in every region, industry sector and stage of development throughout the world. This database contains information on VC transactions, the financed companies and the investors.

Dealroom is a platform that helps investors and companies connect with each other and share data. The platform operates across all investment stages, from seed-stage to late growth-stage. Dealroom enables investors to track companies' progress and decide the appropriate time to invest in them. Entrepreneurs are able to control investment interests and use Dealroom as their official channel for outgoing information to potential investors.

4.1 Access to Dealroom

Access to the Dealroom databases can be obtained via online access through an end-user interface.

The database is not freely accessible but requires an annual subscription. For more information about subscription options we refer to the Dealroom website.

4.2 Dealroom data

Dealroom information up to November 2018 is used for this matching project. The report includes only Venture Capital investments and exits.

4.2.1 Venture Capital investments

Venture Capital investments classified by Dealroom as follows:

- **Angel investment**: Angel investments refer to investments into an early-stage and innovative company made by an Angel Group. An Angel Group is defined as a group of accredited investors that make investment decisions based on the consensus of the membership.
- **Seed Round**: is invested in companies at very early stages of development. It is financing to research, assess and develop an initial concept before a business has reached the start-up phase. Typically, founders and product developers, such as engineers or molecular biologists, are on board, but no complete management team is in place. Most seed rounds do not raise more than $2.5 million.
- **Early, A, B, C, D, E and F**: This ordinal nomenclature is used to describe most venture rounds. Companies often refer to funding rounds as "first," "second," "third," etc. This type of financing is provided to companies for product development and initial marketing. Companies may be in the process of being set up or may have been in business for a short time, but have not sold their product commercially.
- **Later VC:** Later stage financing is provided for the expansion of an operating company, which may or may not be breaking even or trading profitably.

4.2.2 Venture Capital exits

Venture Capital exits reported by Dealroom include VC investments that have been exited through:

- **Acquisition**: A secondary transaction where the purchaser (a company) acquires all of the outstanding equity in a company and, in effect, buys the company.
- **Initial Public Offering (IPO)**: An IPO is an equity financing event where a company raises equity in the public markets for the first time.

4.3 Matching procedure

Records from Dealroom and Innovation Radar databases were matched by mainly using a firm's URL address. As a control check, firm names and location information (country, city and street name) were used.

4.4 Data coverage

Following the matching procedure outlined above, IR organisations could be matched with their corresponding VC information in Dealroom. As VC is mainly provided to startup companies and small businesses, only private firms in IR have been matched with Dealroom.

Out of 1606 private organisations identified by the IR as key innovators in FP projects, 1594 were found in the Dealroom database. Profiles of these companies were enriched with the information available in Dealroom. Information on VC funding or exits was found for 191 or 12 % of all key innovators.

4.5 Data analysis

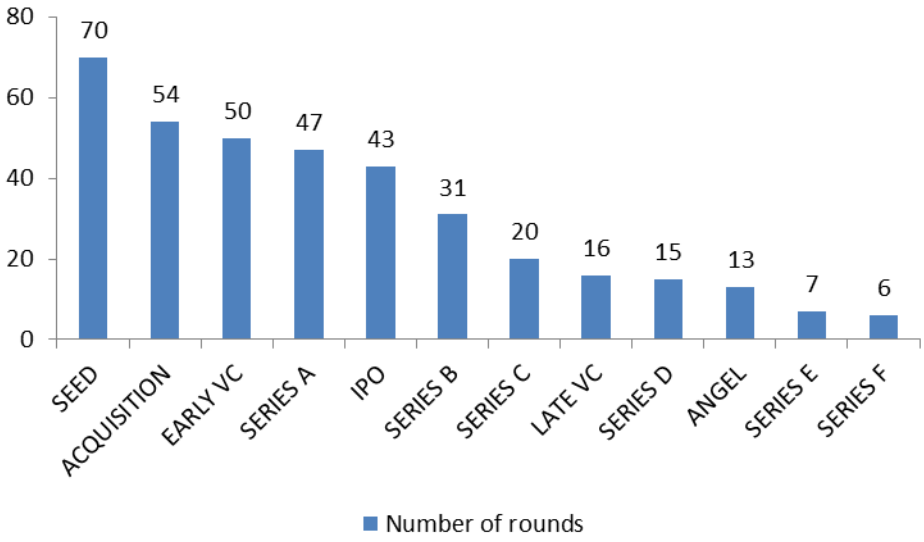
The total number of VC funding or exits, in which the identified firms were involved is 372.

Figure 12 presents the breakdown of the funding rounds by types. Innovators identified by the Innovation Radar were mainly involved in seed and early VC funding rounds. Acquisitions dominate as a form of exit.

Figure 13 presents the distribution of country of origin of matched firms between IR and Dealroom. This country distribution reveals the following:

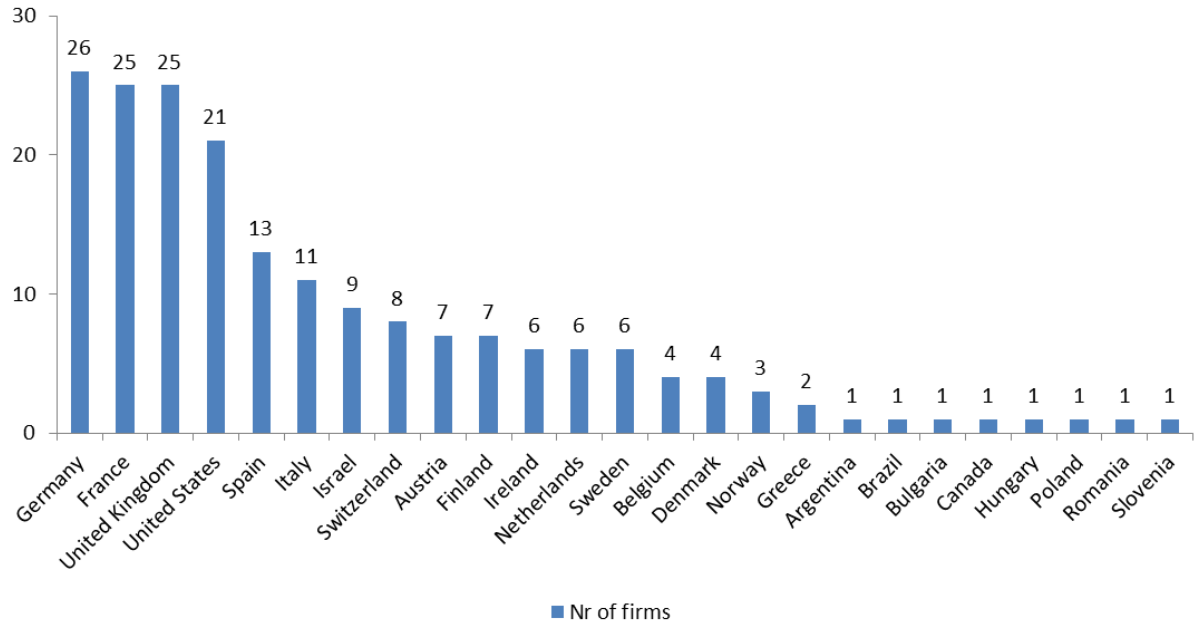
- Altogether, innovators from 25 countries have been found to have received VC capital or were acquired or went public;
- More than half of the matched firms stem from Germany (14%), United Kingdom (13%), France (13%) and the US (11%);
- Spain, Italy, Israel, Switzerland, Austria and Finland have respective percentages around 7-4%;
- The shares of the remaining countries oscillate around 1-3%.

Figure 12: Distribution VC deals and exits among IR key innovators by round type



Note: The figure presents the number of VC funding deals and exits among the innovators identified by the Innovation Radar.
Calculations: JRC

Figure 13: Country of origin of IR key innovators identified in Dealroom



Note: This figure presents the distribution of country of origin of matched firms between IR and Dealroom.
Calculations: JRC

4.6 Alternative VC funding databases

Over the recent years, due to the increase activity in private equity investments, alternative sources of VC funding information have emerged. We provide a non-exhaustive list of alternative databases:

- Venture Source;
- Crunchbase;
- CB insights.

Each of them uses various types of data collection methods such as VC funds, associations, through collection of press releases to self-reporting. A comparison of all data sources and their coverage would require access to each of them. This is beyond the current exercise. One of the advantages of Dealroom over other data sources is its particular emphasis on Europe. Considering that most the companies supported by Framework Programme are based in Europe, this increases the chances of having their funding rounds recorded by Dealroom.

5 Lessons learned

This report presents the process and results of linking the Innovation Radar data with third-party databases to obtain performance information about the organisations that participated in FP projects screened by the Innovation Radar. More concretely, the 5301 organisations in FP projects screened by the Innovation Radar between March 2014 and January 2018 are enriched with:

- Financial information from ORBIS;
- Patent information from PATSTAT;
- Venture Capital funding information from Dealroom.

This section summarises the key lessons learned during this matching process.

Large divergence in access to databases and matching procedures

Linking the Innovation Radar databases to third-party databases is not a straightforward process as each external database provides different access options and requires different matching algorithms. Differentiations in the database access (online platform versus offline content) and internal organisation of the datasets require different downloading procedures and specific computer skills (e.g. knowledge of SQL, fuzzy matching skills).

Database- and topic-specific knowledge is required

The compilation of a new database from external databases requires in first instance a good comprehension of the internal structure of each external database. However – and even more important – it requires an in-depth knowledge and deeper understanding of the topics that are covered in each of these databases. For the current report, in-depth wisdom about accounting statements, company structures, patenting processes and Venture Capital were primordial to understand the internal structure of the respective databases and to successfully manage the matching process.

Data presentation and visualisation is key

The data compilation as presented in this report resulted in a data warehouse filled with raw data. The next challenge will be to put data into a specific context and to provide tools for analysis, aggregation and visualisation in order to present the collected data in a comprehensive way to policy-makers and practitioners.

Profiling of IR participants and hands-on policy support

Enrichment of the Innovation Radar database with other data sources is primordial to increase the quality and depth of the intelligence that can be extracted about innovators and innovations in EU-funded projects. Analyses of the performance of IR participants through financial, patent and Venture Capital funding information aims to facilitate the profilation of IR participants, which can subsequently provide guidance for hands-on policy support initiatives.

References

De Prato, G., Nepelski, D., Piroli, G., 2015. Innovation Radar: Identifying Innovations and Innovators with High Potential in ICT FP7, CIP & H2020 Projects. Joint Research Centre, Science for Policy Report - EUR 27314 EN; doi:10.2791/61591.

EC, 2014. Horizon 2020 Work Programme 2014 - 2015: Leadership in enabling and industrial technologies - Information and communication technologies.

EC, 2018. Recommendation 2003/361/EC: SME Definition. Archived on 2015-02-08. Retrieved from: https://web.archive.org/web/20150208090338/http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/sme-definition/index_en.htm.

Kalemli-Ozcan, S., Fan, J., Penciakova, V., 2015a. Processing ORBIS historical disk. In cooperation with Bureau Van Dijk.

Kalemli-Ozcan, S., Sorensen, B., Villegas-Sanchez, C., Volosovych, V., Yesiltas, S., 2015b. How to construct nationally representative firm level data from the ORBIS global database. National Bureau of Economic Research.

List of figures

List of tables

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: http://europea.eu/contact

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: http://europa.eu/contact

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: http://europa.eu

EU publications
You can download or order free and priced EU publications from EU Bookshop at: http://bookshop.europa.eu. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see http://europa.eu/contact).

KJ-NA-29559-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

Joint Research Centre

EU Science Hub

Publications Office

doi:10.2760/127887

ISBN 978-92-79-98370-2