European Commission

Policy Support Facility Mutual Learning Exercise: Evaluation of Business R&D Grant Schemes

Big data: data linking, new data sources and new data analytics methods

Thematic Report No. 1

[*December 2017*]

Research and Innovation

MLE on the Evaluation of Business R&D Grant Schemes — Big data: data linking, new data sources and new data analytics methods

European Commission
Directorate-General for Research and Innovation
Directorate A: Policy development and Coordination
Unit A4: Analysis and monitoring of national research and innovation policies

Contact (H2020 PSF MLE EVALUATION OF BUSINESS R&D GRANT SCHEMES):
Eva Rueckert, Coordinator of the MLE, Unit A.4 - eva.rueckert@e*c.europa.eu*
Contact (H2020 PSF coordination team):
Román ARJONA, Chief Economist and Head of Unit A4 - Roman.ARJONA-GRACIA@ec.europa.eu
Stéphane VANKALCK, PSF Head of Sector, Unit A4 - Stéphane.VANKALCK@ec.europa.eu
Diana SENCZYSZYN, PSF Team Leader, Unit A4 - Diana.SENCZYSZYN@ec.europa.eu

European Commission
B-1049 Brussels

Manuscript completed in December 2017.

EUROPEAN COMMISSION

Policy Support Facility Mutual Learning Exercise: Evaluation of Business R&D Grant Schemes

Big data: data linking, new data sources and new data analytics methods

Thematic Report No. 1

Prepared by the independent expert: Martijn Poel

Directorate-General for Research and Innovation

2017 EN

Table of Contents

List of tables

List of figures

1 INTRODUCTION

This paper has been prepared for a Mutual Learning Exercise (MLE) on the evaluation of business research and development (R&D) grant schemes in European countries. It addresses the use of big data for the evaluation of business R&D grant schemes (and R&D and innovation policy in general) and the main challenges when doing so. The first section will define and disentangle the concept of big data. Next, we present the objectives and structure of this paper.

1.1 Big data

Around 2010, big data became a popular label for the growing opportunities to collect, process, analyse and use data (Gartner 2011, McKinsey Global Institute 2011, Mayer-Schönberger and Cukier 2013, Kitchin 2014). These opportunities increased thanks to technological progress. First of all, data collection options increased. This is mainly due to cheaper, smaller and better sensors (installed in products and in production systems), greater precision of earth-observation systems (e.g. using satellites for tracking and tracing), more information provided on the internet (enabling web scraping) and more people using online services via the internet or mobile apps (which allows for the monitoring of both potential and actual clients' behaviour).

Data collection, data sharing and data processing have been enabled by improvements in connectivity, data storage and computing power. This concerns servers, computing and other equipment located in companies and other types of organisations as well as advances in cloud computing. For example, companies can collect data from different machines at a production location (e.g. smart factories), to share data between different companies in one value chain (e.g. smart industry) and between offices and mobile colleagues (e.g. sales agents and repair staff). Using on-site or cloud computing, these various data sources can be analysed, visualised and used for decision-making.

Studies about big data emphasise applications by companies and by (smart) cities. Examples are data mining and client profiling by online retailers, online advertisers, banks and insurance companies (McKinsey Global Institute 2011, Mayer-Schönberger and Cukier 2013, Kitchin 2014). Smart city examples include monitoring traffic, air quality and providing navigation suggestions to drivers (Nuaimi et al. 2015). The use of big data by national and international policymakers and public agencies has only recently increased (Technopolis Group, Oxford Institute and CEPS 2015, Bakhshi and Mateos-Garcia 2016).

1.2 Definitions of big data

Volume, variety and velocity of data are at the heart of definitions of big data (Gartner 2011, Mayer-Schönberger and Cukier 2013). In other words, this means larger volumes of data, a greater variety of data sources (including sensor data, transaction data, administrative data, text on websites, surveys, etc.) and high frequency or even real-time collection and processing of data (compared to collection of data once a week, month, etc.). The combination of data variety and velocity often leads to volume (Kitchin and McArdle 2016). One enabler of data volume is that different types of data can be linked, even when there are differences in object identifiers, the timing of measurements, data quality and different categories for recording metadata.

Related to volume, variety and velocity are the possibilities to collect data about 'everything and everyone' (Kitchin and McArdle 2016): in short, from sampling to exhaustivity (n=all). Moreover, data can be collected with little or no structure (e.g. monitoring consumers' online behaviour) or the structure of a dataset can be designed for one purpose but the data may be used for other purposes (e.g. using transaction data for macroeconomic policy). Exhaustivity and alternative uses of data have increased the need for new data analytical methods, to explore whether and how variables interact. Examples are data mining and text mining using machine learning and algorithms in general (Mayer-Schönberger and Cukier 2013, Kitchin 2014).

Veracity, variability, visualisation and value are also mentioned in discussions on big data. Veracity refers to the challenge to ensure data quality in terms of validity and integrity. Variability means that data points allow for the monitoring of *changes* in objects or phenomena of interest. Visualisation is the challenge of presenting larger volumes of data, different types of data, etc. Value refers to the economic and social value of data for the organisation itself and the potential for sharing or selling it (Mayer-Schönberger and Cukier 2013, Kitchin 2014, IDC and Open Evidence 2017).

The definition from Taylor, Schroeder and Meyer (2014, p.1) is compatible with others that focus on the 'Vs' although it also makes it clear that data sources always have a link to objects. Moreover, the definition refers to data analytical tools and to the evolution from small to big data:

> *"Big data is a step change in the scale and scope of the sources of materials (and tools for manipulating these sources) available in relation to a given object of interest."*

This definition has been effective in a study about the state of the art and challenges in the use of big data for policymaking (Technopolis Group et al. 2015).

1.3 Objective and structure of this paper

This challenge paper is prepared in the context of a Mutual Learning Exercise (MLE) on the evaluation of business R&D grant schemes in European countries. Participants in the MLE are R&D and innovation agencies in European and neighbouring countries. The MLE instrument is part of the Policy Support Facility, funded by the European Commission.

Business R&D grant schemes are a specific type of instrument to support companies' R&D and/or innovation. These schemes are provided to individual companies although there can be requirements in terms of collaboration with other companies, universities or research institutes. As such, there can be effects on beneficiaries and their R&D and innovation partners, value chain partners and the regional ecosystem. The *ex-post* evaluation of business R&D grant schemes was the topic of a Mutual Learning Exercise (MLE) that took place in 2016 and 2017 (Cunningham et al. 2017). The main conclusions with respect to evaluation methods were:

- Evaluations using econometric analyses are far from standardised and are quite a complex type of analysis to perform;
- Econometric analyses are very demanding in terms of data availability and quality;
- The working procedures regarding access to data and data confidentiality have only been solved in a few cases;
- There is a trend toward econometric analysis, including the use of control groups, but this needs to be balanced by recognising the simultaneous need to better understand the behavioural effects of using R&D and innovation grants (e.g. the "innovation journey of firms").

Among other things, the conclusions acknowledge the strengths and limitations of quantitative and qualitative methods. For example, it is mentioned that econometric methods can be a 'black box' for non-specialists, including policymakers and politicians (Cunningham et al. 2017).

The present MLE follows on from the 2016-2017 exercise and explores the opportunities and challenges of big data (first workshop), the importance of understanding and measuring behavioural change (second workshop), and recent advances in mixed-method approaches, including econometrics, the use of control groups and qualitative methods (third workshop). All three topics fall within the theme of evaluating business R&D grant schemes.

The first workshop took place on 29-30 August 2017 in Oslo and was hosted by Innovation Norway. It addressed data linking and the use of new data sources, new data analytical methods and tools. The workshop also addressed challenges concerning data platforms (including technical facilities), data sharing, data quality and privacy. The draft challenge paper helped participants to prepare for the workshop. The challenge paper benefited from the MLE kick-off meeting (9 June 2017 in Brussels) and a short survey among the MLE participants.

Section 2 will discuss the use of big data for R&D and innovation policy, such as the various schemes/instruments used to support R&D and innovation.

Section 3 will zoom in on the use of big data (especially data linking) for the evaluation of business R&D grant schemes.

Section 4 will address methodological challenges as well as data sharing, ethical and related challenges. It will also touch on the added value of using big data to evaluate business R&D grant schemes.

2 THE USE OF BIG DATA FOR R&D AND INNOVATION POLICY

2.1 Introduction

The use of big data for the design and evaluation of R&D and innovation policy is increasing. This can be seen most clearly for R&D and innovation activities performed by universities, research institutes, individual researchers and their (social) networks. In line with the discussion above, there has been an evolution. For instance, bibliometric and patent databases have been used for decades and continue to be very useful. To a large extent, the increased use of big data for R&D and innovation policy is discussed under the altmetrics label (Galligan and Dyas-Correia 2013, Haustein et al. 2014).

Below, we address data linking ('from data variety to data volume'), the use of new data sources (such as data available on the internet) and using new data analytical methods and tools. Next, we discuss two enablers for using big data: technical platforms and shared ontologies.

2.2 Data linking

Data linking is used in evaluations of various types of R&D and innovation support schemes. Data linking can be observed in monitoring schemes and evaluations of research-industry collaboration. An example is having unique identifiers of universities and companies and exploring collaborations in R&D projects, whilst also looking at co-publications and joint patents (Fraunhofer ISI et al. 2009, Gal et al. 2016).

Moreover, data linking is used in evaluations of R&D programmes, research institutes and specific support schemes such as fiscal incentives (Technopolis Group and SEO 2016a, Research Council of Norway 2016, Hertog et al. 2016, respectively). Here, we also see applications that address *companies*. Datasets include company databases and official statistics to explore how companies evolve (in terms of employment, revenues, exports, etc.), innovation surveys, patent databases and other data sources to explore whether and how these companies innovate, and data about companies using support schemes (often involving control group approaches).

Data linking can also be observed in university rankings. For example, the Times Higher Education World University Rankings links data about publications, citations, reputation, research income, staff/student ratio and other indicators.

2.3 New data sources

Using new data sources mainly concerns the use of data sources available on the internet. The wealth of such data and information provides the best illustration of data variety, volume and velocity. Examples include publications and citations on Google Scholar (Harle et al. 2016, Prins et al. 2016), the visibility of research results on social media such as Twitter and Facebook (Thelwall et al. 2013, Ringelhan et al. 2015) and references to research themes, research articles and reports in policy documents (Bornmann et al. 2017).

To some extent, studies are linking new data sources to established data sources. Examples are using Google Scholar publications in university rankings (Daraio and Bonaccorsi 2017) and using social media data to predict whether articles will be cited in academic articles (Peoples et al. 2016).

2.4 New data analytical methods

New analytical methods and tools are applied to new data sources, such as social media data, documents available on the internet and (open) databases and repositories like Google Scholar and the datasets in the EU Open Data Portal. Web scraping is one of the more recent tools, complemented by text mining (Thelwall et al. 2013, Gök et al. 2015).

Text-mining applications can be relatively straightforward, using sets of keywords, but there are also examples of machine learning for exploring relevant concepts, relevant relations between actors, and relevant geographic locations for specific technologies and industries. Examples include text mining of patent databases to explore technology trends (Balsmeier et al. 2016) and text mining of business registries to measure industries that are hard to capture using static industry classifications (Bakhshi and Mateos-Garcia 2016).

Moreover, the availability of more data linking opportunities and greater data volumes increases the potential of econometric methods. For example, the availability of more data increases the possibility of creating control groups, adding control variables and assessing the impact of the main explanatory variables. In the context of big data, experts have stressed the relevance of econometric methods, such as Bayesian techniques, Rubin's Causal Model and Propensity Score Matching (Varian 2014, UK Department for Business, Energy and Industrial Strategy 2017).

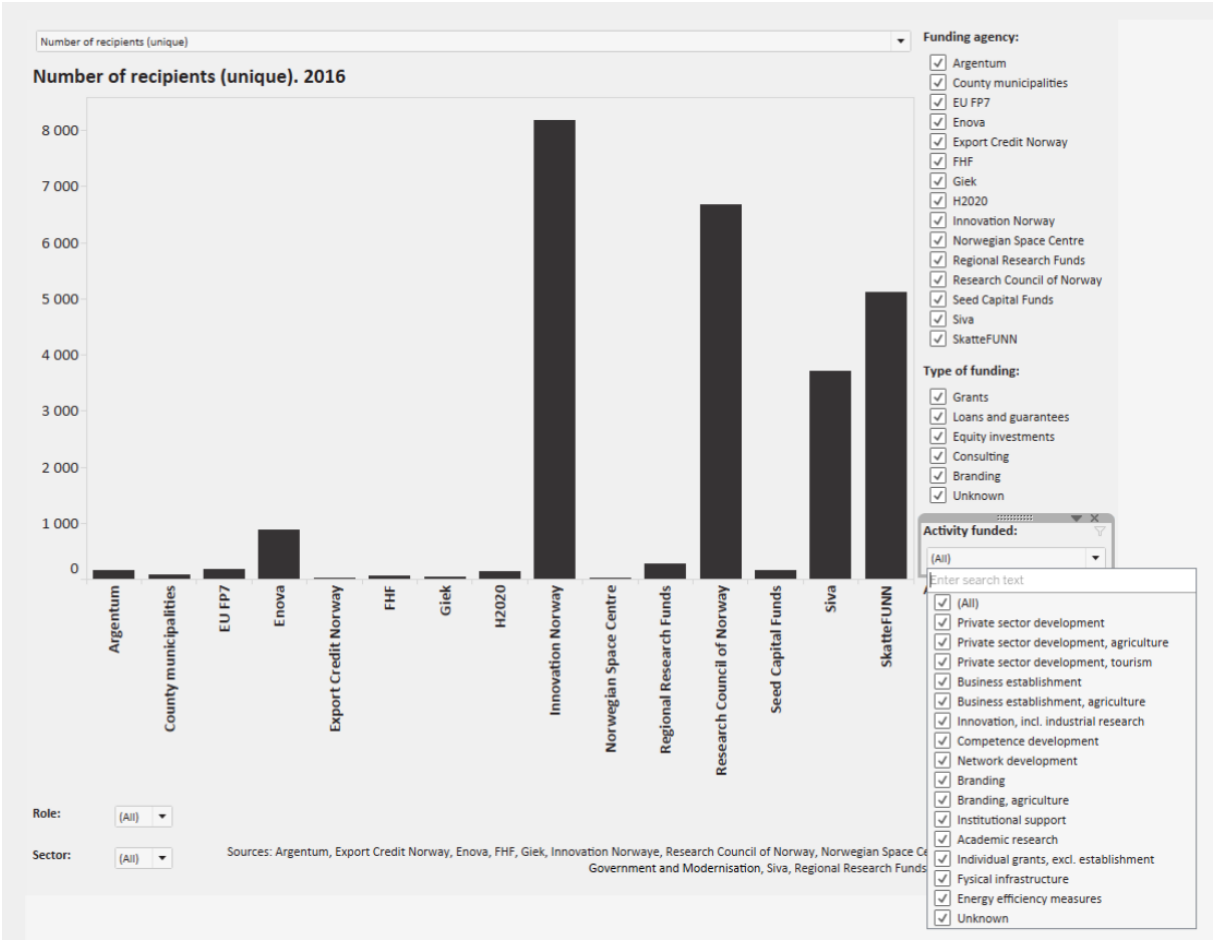
2.5 Technical platforms and shared ontologies

Important requirements for using big data for R&D and innovation policy are: 1) technical platforms for storing, linking and sharing data; and 2) shared ontologies for different types of R&D and innovation, different types of organisations, different R&D and innovation support schemes, and different types of effects and other variables (and corresponding indicators).

Innovation agencies which invested substantially in technical platforms include Innovate UK (the innovation funding service), Innovation Norway (linking data from different agencies) and VINNOVA Sweden (investing in data linking and open data).

Figure 1 shows that Innovate Norway, other Norwegian funding agencies and the Research Council of Norway share their data on organisations which received support for R&D and innovation. Using this integrated and shared data platform, agencies and analysts in Norway can analyse which organisations receive multiple types of public support and how this changes over time, as companies grow and/or their R&D and innovation activities change. The data in Figure 1 is for illustration purposes only because relevant agencies are in the process of uploading their data.

Figure 1: Norwegian data platform for monitoring which organisations receive support from which agencies and support schemes



Source: Innovation Norway and Norwegian Ministry of Trade and Fisheries

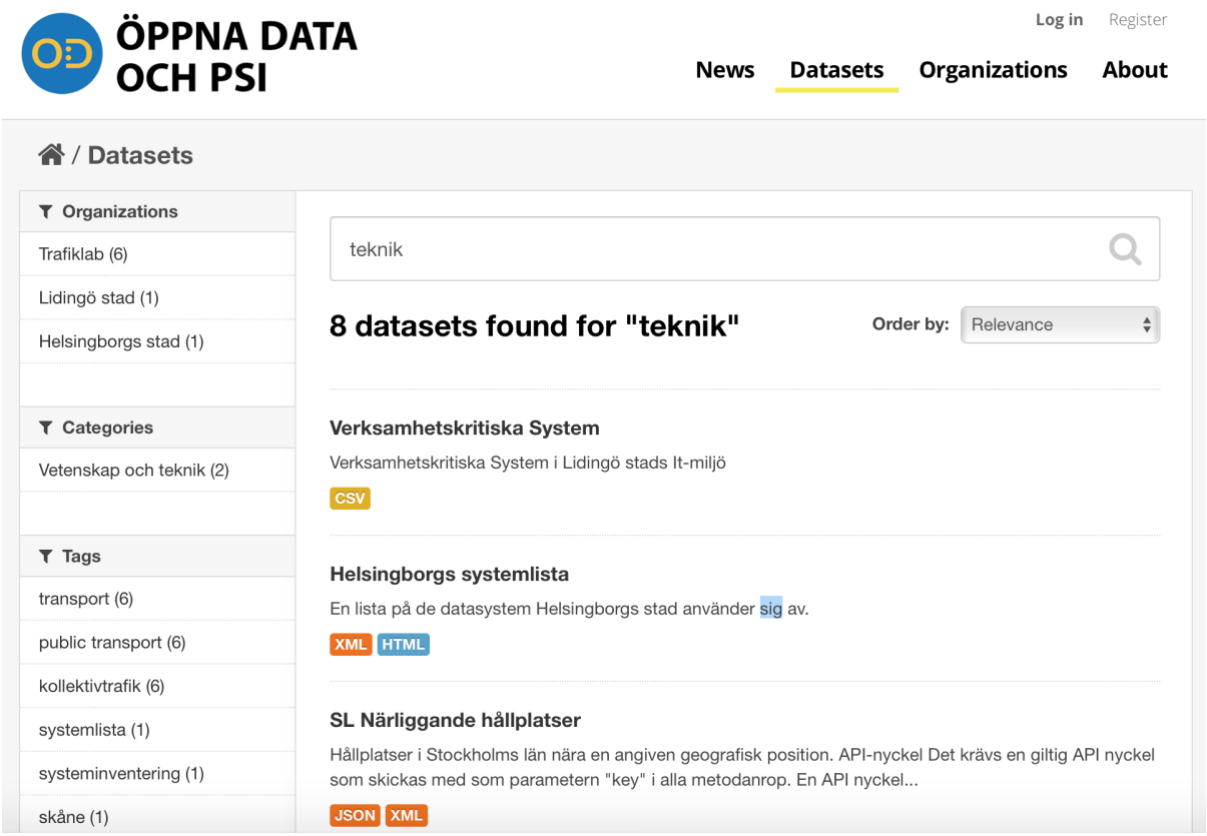
For Innovation Norway, this joined-up initiative will complement an internal initiative which links the data on all of Innovation Norway's support schemes. Staff members have easy access (using Microsoft's PowerBI dashboard/front-end) to data about organisations that receive(d) support from Innovation Norway. This allows for basic checks and descriptive statistics as well as for exploring how Innovation Norway supports companies across their 'innovation journey'.

Another example is the UK Gateway to Research system that was initiated by Research Councils UK and Innovate UK. Although this system emphasises sharing information about research projects (activities and results), it can also be used to analyse which organisations receive which types of public support. Data about more support schemes is gradually added to the system and more data is disclosed in full online (open data)[1].

Figure 2 illustrates how Vinnova, the Swedish innovation agency, and Swedish partners present their data about innovation projects, business locations, traffic, houses, population density, etc. One of the many applications links data about companies, e.g. whether companies that are newly created and/or benefit from R&D and innovation support influence the development of specific cities and regions.

1 See for example the descriptive analysis in the Innovate UK 2015/2016 funding report: https://innovateuk.blog.gov.uk/2017/08/29/innovate-uks-201617-funding-reports-what-do-they-tell-us/

Figure 2: Vinnova open data platform



Source: Open Data Sweden and Public Service Innovation: https://oppnadata.se/en

In some countries, ministries are playing a significant role in developing technical platforms for storing, processing and sharing data on R&D and innovation. An example is the database of the Centre for the Development of Industrial Technology, a Spanish public organisation under the Ministry of Science and Innovation. Other examples are the open data and linked data platform of the Welsh government[2] and the Entrepreneur Innovation System (EIS) of the Turkish Ministry of Science, Industry and Technology (MoSIT).

Organisations investing in shared ontologies include US federal agencies (the Star Metrics project), the European Commission and the Organisation for Economic Co-operation and Development (OECD) (the Innovation Policy Platform, the REITER platform and the STIP project: International Survey on Science, Technology and Innovation Policies).

The REITER initiative, coordinated by the OECD, will develop an ontology of policy instruments for supporting R&D and innovation. This ontology will be used to structure a survey among national policymakers and to analyse the results across OECD countries. The OECD also plans to share the full dataset ('which policy instruments, for which target groups, in which countries') as an online and searchable database[3].

2.6 From academic and explorative studies to policy evaluations

The use of big data for R&D and innovation policy mainly concerns academic research and explorative studies rather than commissioned research and policy evaluations. However, data linking is also used in policy evaluations. This picture is fully consistent with an analysis of 58 initiatives that use big data for policymaking, in the field of R&D innovation,

2 Available in beta version: http://gov.wales/about/foi/open-data/?lang=en

3 More information can be found at: https://www.innovationpolicyplatform.org/stip-monitoring-and-analysis-ec-oecd-project/semantic-technologies-and-semantic-web-structuring-data

environmental policy and other policy fields. A few evaluation studies use data linking and, in particular, new data sources and new methods and tools (Technopolis Group et al. 2015).

In other words: the use of new data sources like social media and website data and new tools such as web scraping and text mining are not yet mature enough for evaluation purposes. A clear and realistic assessment of the level of maturity is ever more important, given the high stakes (e.g. adapting or stopping support schemes) and the persistent challenges of evaluating R&D and innovation support schemes. Evaluation challenges include: skewed impact distributions; the time lag between policy support and outcomes; difficulties in attributing outcomes and changes in the behaviour of organisations towards one support scheme; and communicating the results of evaluations to policymakers and politicians (as discussed in the previous MLE: Cunningham et al. 2017).

Among other things, the use of new data sources has the potential to address the skewed impact distribution of R&D and innovation projects (few projects, companies or partnerships achieving substantial outcomes). Text mining enables the analysis of a large number of projects, deliverables, company websites, etc. which reduces the chances of successful projects being overlooked.

3 THE USE OF BIG DATA FOR THE EVALUATION OF BUSINESS R&D GRANT SCHEMES

3.1 Introduction

As mentioned in section 1, the evaluation of business R&D grant schemes was the topic of the MLE that preceded the current one. Because business R&D grants are one specific type of scheme among many others under the umbrella of R&D and innovation policy, it is challenging to identify evaluations of business R&D grants which use big data. We have identified a few examples, and refer to other types of support schemes that target companies.

Below, we address data linking, new data sources and new data analytical methods. However, we do not discuss technical platforms and shared ontologies (see section 2) as these are seldom sufficiently specific for business R&D grant schemes.

3.2 Data linking

Although the 2016-2017 MLE did not address big data, it was mentioned as relevant for improving evaluation approaches. Data linking was considered as one way to improve the richness of those evaluation studies that apply econometric methods (Cunningham et al. 2017). One example given is the evaluation of a voucher scheme. The dataset of companies that received a business grant (and those that applied for a voucher but did not receive one) was linked to a commercial database with financial data about companies (the beneficiaries and the control group). The voucher scheme increased the productivity of beneficiaries, especially micro-firms (Christensen et al. 2015). However, econometric approaches, with or without data linking, are more common in *academic* studies than in commissioned evaluations of business R&D grants. Official evaluations tend to address the rationale of the intervention and implementation performance, including output (Cunningham et al. 2013).

The What Works Centre for Local Economic Growth (2015) provides an indication that the opportunities to use data linking are seldom applied to evaluating business R&D grant schemes. This UK-based research centre reviewed the datasets and methodologies used in evaluations of business grants, loans and subsidies. One of its conclusions is that evaluations tend to focus on a small number of outputs and outcomes, while only using the corresponding datasets for these outputs and outcomes (or relying on surveys of beneficiaries). As such, there is limited use of additional datasets to explore long-term effects on beneficiaries, e.g. the process or journey from R&D to innovation, revenue growth, more export, etc. Similarly, there is no information about the effects of supporting individual companies on their value chain partners and business ecosystem.

> *"As discussed above, relatively few studies consider more than one element of the chain from increased R&D spend, through innovation, to improved firm performance. The one evaluation (study 467) that looks at all three elements finds no effect on R&D spend, and no effects on patents or product innovation. It does, however, find positive effects on self-reported process innovation. Somewhat puzzlingly, this does not show up in increases in productivity where the study finds zero effects. It does, however, find weakly positive effects on employment, sales growth and exports.*
>
> *Among the five studies that look at both innovation and economic outcomes, only one finds consistently positive effects on both. The second finds positive effects on patents, but mixed effects on employment and no effect on profits. The third reports that R&D subsidies had a positive effect on employment but no effect on patents. A fourth finds no effect across all outcome variables considered: patents, employment, productivity and sales. A fifth finds no effect on patents, but positive effects on self-reported innovation and on sales due to new products/services." (What Works Centre for Local Economic Growth 2015, p.29).*

A small survey among the innovation agencies that have participated in the MLE (10 responses) indicates that half of them have monitoring strategies (and technical platforms) that link different datasets. Other innovation agencies consider data linking a task for partners, such as statistical offices, research councils, research institutes (e.g. ZEW in Germany) and consultants. As such, innovation agencies invest in linked data that can be used in future evaluations of business R&D grant schemes.

Table 1 provides examples of evaluations of R&D and innovation support schemes targeting businesses and (in two case) other organisations too. In line with the observations made above, it was a challenge to identify evaluations of business R&D grant schemes that applied data linking.

The table illustrates the types of datasets that can be linked as well as the importance of having unique identifiers for companies. Table 1 also illustrates that the use of control groups has become the norm when using quantitative methods to evaluate support schemes that target businesses (and other organisations).

In the small survey among innovation agencies, similar types of datasets were mentioned, as were unique identifiers such as tax codes and company names. It was also mentioned that even with unique identifiers (and especially when using company names), removing double entries, merging data, etc. must be done manually.

Table 1: Data linking: examples of evaluations of R&D and innovation support instruments

Study	Ministry/ agency	Datasets linked	Unique identifier	Method used
BEIS (2017), The impact of public support for innovation on firm outcomes	UK Department for Business, Energy & Industrial Strategy	Administrative data of Innovate UK and the National Measurement System (firms that sought advice or support for research or measurement services), Business Structure Database, the Business Enterprise R&D dataset	IDBR enterprise reference number and the Companies House Reference Numbers (CRN) that are equivalent to enterprise reference numbers from the Office of National Statistics	Propensity score matching, descriptive analysis
Dialogic (2017) Evaluation of the SBIR instrument	Dutch Ministry of Economic Affairs	Business register, R&D expenditure (WBSO database), revenue and profit data	Trade registers number, business ID from business register	Difference-in-difference, regression discontinuity analysis
Technopolis Group (2016a), Evaluation of the Dutch Association for Technical Sciences (STW)	Dutch Ministry of Economics and NWO	Web of Science (publications, citations), PATSTAT, NWO database	Company name	Difference-in-difference, regression discontinuity analysis, fixed-effects analysis
Technopolis (2016b), *Ex-post* evaluation of Ireland's participation in the 7th EU Framework Programme	Department of Jobs, Enterprise and Innovation Ireland	CORDA, Annual Business Survey of Economic Impact (ABSEI), Annual Employment Survey (AES), SESAM/RESPIR	ABSEI code	Descriptive analysis, exploratory analysis of trends, Probit analysis
Hertog et al. (2015), Evaluation of the Dutch Innovation Box 2010-2012	Dutch Ministry of Finance	Business register, Corporation tax, Innovations surveys (CIS, RTD), R&D tax credit data	Trade registers number, fiscal code, business and person ID from business register	Difference-in-difference, propensity score matching, first differences
Merito et al. (2010), Do incentives to industrial R&D enhance research productivity and firm growth? Evidence from the Italian case	Academic study, in collaboration with the Italian Ministry of Research	Database of the Italian Ministry of Research, company database (Amadeus/Bureau van Dijk), patent databases (Delphion/ Thompson)	Company name	Stratified random sampling (control group)

A closer look at the six evaluations reveals that the datasets are used to provide data about beneficiaries and, in four cases, a control group. The datasets mentioned, and data linking in general, are seldom used to explore the wider impact of business R&D grant schemes. Examples could be the impact of business R&D grant schemes on value chain partners and regional clusters.

3.3 New data sources and new data analytical methods

According to our literature review of big data and evaluation methods in the field of R&D and innovation policy, new data sources are seldom used to evaluate business R&D grant schemes. The same picture emerges from the small survey among innovation agencies. The 10 survey responses include two examples: first, using web scraping to explore which companies are innovating (e.g. in specific technologies or products). However, this is a methodological experiment and an explorative study rather than an official evaluation. Second, using text mining to analyse the beneficiaries' final reports. This allows for an exploration of a range of outputs and outcomes, expected and unexpected, related or unrelated to each other ('patterns').

Another example was mentioned in an earlier study (Technopolis Group et al. 2015). To evaluate Spanish schemes supporting ICT companies, traditional datasets were linked to data from job-search websites (do beneficiaries of support schemes recruit new employees?) and online media (are beneficiaries mentioned?).

Moreover, text mining of company websites and business registries in the UK is used to identify 'regional hot spots' in industries that hardly fit industry classifications such as NACE. One example is a study of the gaming industry (Bakhshi and Mateos-Garcia 2016). However, this explorative analysis was done in the context of innovation policy in general, rather than being used to design or evaluate business R&D grant schemes (e.g. improving procedures to identify and approach companies in high-priority industries).

Lastly, text mining is used in an evaluation of a UK programme to support engineering research and postgraduate training (Technopolis Group 2015). Companies are among the main targets groups of this programme. As part of the monitoring and evaluation cycle, programme beneficiaries provide quantitative data as well as case studies. Text mining is applied to identify which companies (irrespective of their official classification) and which research organisations are active in emerging fields such as high-performance computing. As such, this is another example that is not 'spot on' in terms of using big data for the evaluation of business R&D grants.

4 CHALLENGES

4.1 The added value of using big data for the evaluation of business R&D grant schemes

The preceding chapters have discussed how data linking is used more frequently than new data sources when evaluating business R&D grants schemes. The same applies to other schemes under the R&D and innovation policy umbrella. As a result, there is continued interest in using econometric methods to analyse linked and structured datasets. Because new data sources, such as social media data, company websites and job-search websites, are only used to a certain extent, few examples of new data analytical methods and tools have been mentioned. Examples could include machine learning to explore why, how and which companies invest in R&D (and achieve results), to identify relevant types of organisations in social networks ('profiling'), and perceptions about technologies, research topics or companies ('sentiment mining'). Text mining can also be applied, more often, to extract relevant data from commercial databases and business registries.

- Innovation agencies, policymakers and other stakeholders should continue to experiment with big data to assess the added value compared to relying on traditional approaches alone. For example, how can web data help to:

- monitor and analyse a larger range of companies (beneficiaries of both single and multiple schemes as well as control groups);

- address specific steps in the process/journey of R&D, innovation and business performance (e.g. recruiting staff, increasing visibility, reputation and internationalisation);

- collect data at shorter intervals (e.g. compared to official statistics); and

- triangulate the results achieved with other methods (e.g. self-reporting in surveys and progress reports)?

4.2 Methodological challenges

Having or developing unique identifiers is a requirement for data linking. As mentioned in section 3, several options are being used. The process of data linking should be transparent, e.g. why specific companies or data points were excluded from the linked dataset. More generally, transparency is a requirement for big data approaches, as it is for using single datasets and established methods – in short: no black boxing (Pasquale 2015).

Transparency does not only concern the methodological details, such as the algorithms, and the extent to which the underlying data is open data. It can also be enhanced by making explicit the intervention logic of a support scheme as a basis for developing indicators and collecting data. This avoids situations in which data availability steers the design of evaluation studies strongly towards data that is easily available rather than data that is most relevant. Another pitfall is relying too much on stakeholders to collect data, e.g. using self-reported and confidential data, without any possibilities for triangulation. This could lead to policy-based evidence rather than evidence-based policy (Strassheim and Kettunen 2014).

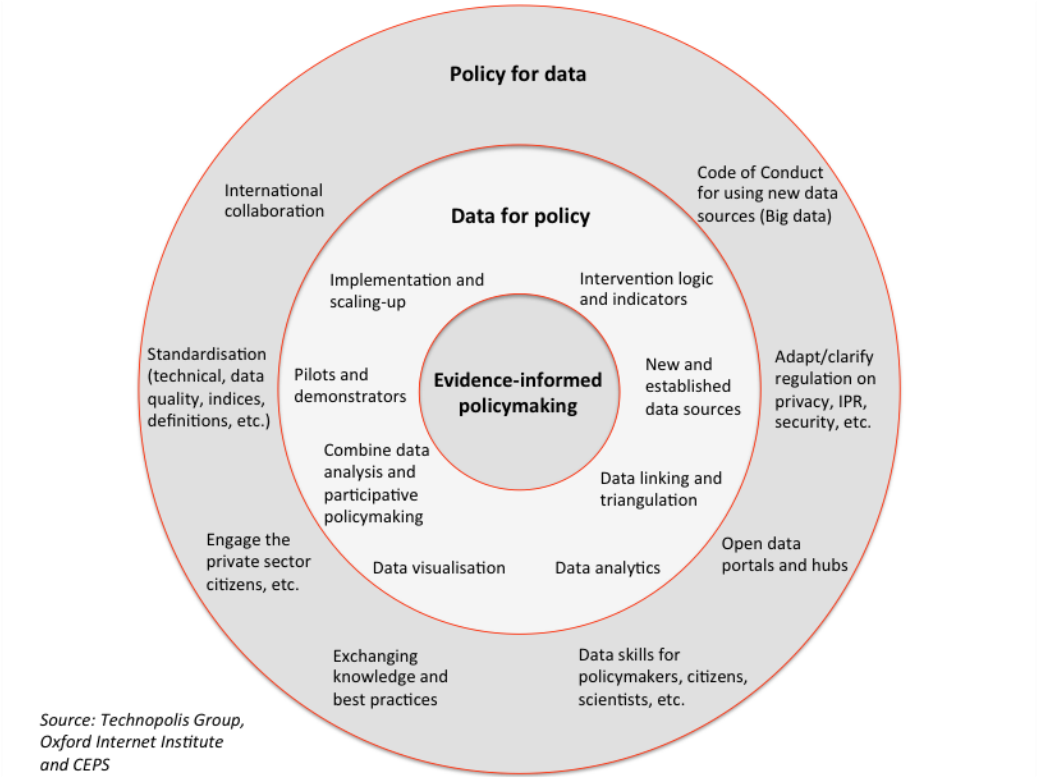
One methodological, or rather, conceptual challenge is to develop ontologies that are shared among relevant agencies in one country and, if possible, with international peers and organisations such as the OECD and the European Commission. This concerns the ontology for types of support schemes as well as types of outputs and outcomes, and types of organisations. At some point, linked datasets can be used to evaluate a range of support schemes (and the policy mix) on a number of indicators, for different types of organisations.

As mentioned in section 2, other challenges concern the technical platform for data storage, data linking and data sharing. Four specific points are the versatility (agility) and scalability of the platform, the extent to which one or several agencies manage it, and the dependency on IT system suppliers.

4.3 Data sharing, ethical and related challenges

As part of a 2015-2016 study on the use of big data for policymaking (Technopolis Group et al. 2015), an expert workshop was organised to identify the main challenges. Rather than discussing 'data for policy', the experts preferred to address 'policy for data'. Figure 3 summarises the types of policies and framework conditions that should be in place for the effective and ethical use of big data.

Figure 3: Data for policy and policy for data



The challenge of standardising definitions has already been mentioned in section 2. Shared ontologies for the field of R&D and innovation (including the types of support schemes) facilitate national and international collaboration. One example is collaboration between different national agencies active in the wider field of R&D, innovation, industrial and labour market policies. How do the various types of schemes (the policy mix) influence the various types of companies?

Innovation agencies, like statistical offices and academics, could engage in discussions on codes of conduct for using big data. These discussions about ethics already take place to some extent at the national level (e.g. Responsible Data Analytics in Australia), European level (e.g. the European Statistical System) and at the global level (e.g. the Data Science Association and the United Nations Development Group).

One of the codes of conduct topics is data ownership, including permission for sharing data and the privacy regulations to be respected. A related topic is public-private collaboration, e.g. using commercial databases for policy analysis. Tensions can arise between delivering

a transparent analysis and not disclosing (high-value) commercial data that is owned by private companies.

Another code of conduct issue is inclusion. Using new data sources may imply that some 'objects', such as types of companies, citizens or regions, are covered less than others. For instance, using web scraping in policy evaluations may discriminate against companies with limited or no web presence.

Moreover, innovation agencies and policymakers in the field of R&D and innovation policy can continue to invest in open data. Again, data confidentiality and the privacy of individuals can pose a challenge. On the other hand, open data portals allow evaluators and academic researchers to evaluate the effectiveness of support schemes and to explore trends in funding emerging technologies and industries.

Having sufficient data skills is another issue, which concerns researchers, consultants and policy analysts (e.g. at innovation agencies) and their skills in applying new methods and tools. However, it also concerns senior management, policymakers, politicians and other stakeholders and their skills in interpreting the results of big-data-based studies. Note that advances in visualisation, such as online interactive dashboards, can improve the communication of evaluation results to policymakers and politicians. However, this has to be balanced against the dangers of oversimplification of the data and the underlying assumptions of the analyses.

Exchanging knowledge and best practices is a challenge because many innovation agencies and policymakers have only just started to experiment with big data. This implies that few best practices can be shared, although relevant information about plans and first attempts can be shared. Our MLE on the evaluation of business R&D grant schemes in European countries is one such opportunity for sharing this information.

REFERENCES

Bakhshi, H. and Mateos-Garcia, J. (2016). *New Data for Innovation Policy*. Nesta Working Paper.

Balsmeier, B., Li, G.C., Chesebro, T., Zang, G., Fierro, G., Johnson, K. and Fleming, L. (2016). Machine learning and natural language processing on the patent corpus: data, tools, and new measures. *UC Berkeley Research Paper*, November 2016.

Bornmann, L., Haunschild R. and Marx, W. (2017 *forthcoming*). Policy documents as sources for measuring societal impact: How often is climate change research mentioned in policy-related documents?

Christensen, T.A., Kuhn, J.M. Schneider, C. and Sørensend, A. (2016). Science and Productivity. Evidence from a randomized natural experiment. Prepared for the *SIMPATIC* final conference, Brussels, February 2015.

Cunningham, P., Gök, A. and Laredo, P. (2013). *The impact of direct support to R&D and innovation in firms*. Nesta Working Paper 13/03.

Cunningham P., Hertog P. den and Peter, V. (2017). Ex-post *evaluation of business R&I grant schemes*. Expert panel report prepared for the Horizon 2020 Policy Support Facility, Mutual Learning Exercise. Brussels: European Commission.

Daraio, C. and Bonaccorsi, A. (2017). Beyond university rankings? Generating new indicators on universities by linking data in open platforms. *Journal of the Association for Information Science and Technology*, *68*(2), 508-529.

Dialogic (2017). *Evaluation of the SBIR instrument*. Report for the Dutch Ministry of Economic Affairs. Utrecht: Dialogic.

Fraunhofer ISI, Idea Consult and SPRU (2009). *The Impact of Collaboration on Europe's Scientific and Technological Performance*. Final Report, Karlsruhe, Brussels, Brighton, March 2009.

Gal, D., Glanzel, W. and Sipido, K.R. (2016). Mapping cross-border collaboration and communication in cardiovascular research from 1992 to 2012. *European Heart Journal,* 2016-0, 1-10.

Galligan, F. and Dyas-Correia, S. (2013). Altmetrics: Rethinking the way we measure. *Serials review*, *39*(1), 56-61.

Gartner (2011). *Information management in the 21st century*.

Gök, A., Waterworth, A. and Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics* 102, 653-671.

Harle, C.A., Vest, J.R. and Menachemi, N. (2016). Using Bibliometric Big Data to Analyze Faculty Research Productivity in Health Policy and Management. *Journal of Health Administration Education*, 33(2), 285-293.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H. and Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics*, 101(2), 1145-1163.

Hertog, den P., Vankan, A., Janssen, M., Minne, B., Korlaar, L., Erven, B., Verspagen, B. and Mohnen, P. (2015). *Evaluation Innovation box 2010-2012* (in Dutch). Report for the Dutch Ministry of Finance. Utrecht: Dialogic.

IDC and Open Evidence (2017). *European Data Market*. Brussels: European Commission.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Los Angeles: Sage.

Kitchin, R. and McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society* 3(1): 1-10.

Mayer-Schönberger V. and Cukier, K. (2013). *Big Data: A Revolution that Will Transform How We Live, Work, and Think*. New York: Houghton Mifflin Harcourt.

McKinsey Global Institute (2011). *Big data: The next frontier for innovation, competition, and productivity*.

Merito, M., Giannangeli, S. and Bonaccorsi, A. (2010). Do incentives to industrial R&D enhance research productivity and firm growth? Evidence from the Italian case. *International Journal of Technology Management*, 49(1-3), 25-48.

Nesta (2016). *Innovation Analytics: A Guide to New Data and Measurement in Innovation Policy*. London: Nesta.

Al Nuaimi, E., Al Neyadi, H., Mohamed, N. and Al-Jaroodi, J. (2015). Applications of big data to smart cities. *Journal of Internet Services and Applications*, 6(1), 25.

OECD (2006). *Government R&D Funding and Company Behaviour: Measuring Behavioural Additionality*. Paris: OECD.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Cambridge, Massachusetts: Harvard University Press.

Peoples, B.K., Midway, S.R., Sackett, D., Lynch, A. and Cooney, P.B. (2016). Twitter predicts citation rates of ecological research. *PloS one*, 11(11), e0166570.

Prins, A., Costas, R., van Leeuwen, T.N. and Wouters, P.F. (2016). Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. *Research Evaluation*, 25(3), 264-270.

Research Council of Norway (2016). *Evaluation of Norwegian Technical Industrial Research Institutes*.

Ringelhan, S., Wollersheim, J. and Welpe, I.M. (2015). I like, I cite? Do Facebook likes predict the impact of scientific work?. *PLoS One*, 10(8), e0134389.

Strassheim, H. and Kettunen, P. (2014). When does evidence-based policy turn into policy-based evidence? Configurations, contexts and mechanisms. *Evidence & Policy: A Journal of Research, Debate and Practice* 10(2): 259-277.

Taylor, L., Schroeder, R. and Meyer, E. (2014). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? *Big Data & Society,* July-December 2014, 1-10.

Technopolis Group (2014). *Evaluation Reference Model*. Study Prepared For TAFTIE's Taskforce Benchmarking Impact, Effectiveness and Efficiency of Innovation Instruments.

Technopolis Group (2015). *Assessing the economic returns of engineering research and postgraduate training in the UK*. Study prepared for the UK EPSRC and the Royal Academy of Engineering.

Technopolis Group, Oxford Internet Institute and CEPS (2015). *Data for Policy: A study of Big Data and Other Innovative Data-Driven Approaches for Evidence-Informed Policymaking: Report about the State-of-the-Art*. Prepared for the European Commission.

Technopolis Group (2016a). *Evaluatie Stichting voor de Technische Wetenschappen* (STW) (in Dutch). Report for the Dutch Ministry of Economic Affairs and NWO.

Technopolis Group (2016b). *Ex-post evaluation of Ireland's Participation in the 7th EU Framework Programme*. Report for the Department of Jobs, Enterprise and Innovation in Ireland (DJEI).

Thelwall, M., Haustein, S., Larivière, V. and Sugimoto, C.R. (2013). Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLOS ONE* 8.

UK Department for Business, Energy & Industrial Strategy (2017). *The impact of public support for innovation on firm outcomes*. BEIS Research Paper Number 3. London: BEIS.

United Nations Development Group (2017). *Guidance Note on Big Data for Achievement of the 2030 Agenda: Data Privacy, Ethics and Protection*. New York: UN.

Varian, H.R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2), 3-27.

What Works Centre for Local Economic Growth (2015*). Innovation: grants, loans and subsidies*. Evidence Review 9. London: What Works Centre for Local Economic Growth.

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://europa.eu/contact

ON THE PHONE OR BY E-MAIL
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)

EU LAW AND RELATED DOCUMENTS
For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

OPEN DATA FROM THE EU
The EU Open Data Portal (http://data.europa.eu/euodp/en/data) provides access to datasets from the EU. Data can be downloaded and reused for free, both for commercial and non-commercial purposes.

This document is prepared for a Mutual Learning Exercise about the evaluation of business R&D grant schemes. The paper discusses the use of big data to evaluate business R&D grant schemes and other types of support for R&D and innovation. One aspect of big data – data linking – is being implemented by several public agencies. Other aspects of big data, such as web scraping, text mining and machine learning, are less mature. The paper provides examples from Norway, Spain, Sweden, the Netherlands, the UK and other countries. The challenges of using big data are not only related to data collection, data platforms and data analytics. Equally challenging is the development of shared ontologies of relevant R&D actors, activities, support schemes and results. Such a shared conceptual framework is essential for linking data from different sources. Data confidentiality and inclusion are two important ethical challenges.

Studies and reports