Improving the measurements of innovation outcome

Project : ESTAT/G/2015/006
Working paper 3
Document code: Dialogic v14
Submission date: 30 August 2016

Content:

ACRONYMS

CEO	Chief Executive Officer
CFO	Chief Financial Officer
CIS	Community Innovation Survey
CTO	Chief Technology Officer
EGR	EuroGroups Register
ESS	European Statistical System
FTE	Full-time equivalent
FRIBS	Framework Regulation Integrating Business Statistics
IT	Information Technology
MIP	Mannheim Innovation Panel
NACE	Nomenclature statistique des activités économiques dans la Communauté Européenne
NESTI	OECD Working Party of National Experts on Science and Technology Indicators
NSO	National Statistical Office
OM	Oslo Manual
OM2005	Oslo Manual, 2005 (2^{nd} edition)
PIN	Personal Identification Number
R&D	Research & Development
RFID	Radio-frequency identification
SBRL	Standard Business Reporting Language
SMDX	Statistical Data and Metadata eXchange
VAT	Value Added Tax
UUID	Universally Unique Identifier
XBRL	eXtensible Business Reporting Language

1. Background

1.1 The Oslo Manual as a framework for innovation statistics

The Oslo Manual provides guidelines for developing internationally comparable innovation indicators. In fact, several innovation business surveys at international level – including the Community Innovation Survey (hereafter: CIS), launched in parallel with the first edition of the Manual - largely follow the definitions and statistical practices recommended by the Oslo Manual. Changes in the Manual are expected to percolate through many national and international measurement instruments, including obviously CIS. Likewise, lessons drawn from the use of CIS and other relevant measurement instruments concerning, for instance, the scope, depth, quality and validity of data, are expected to fed back into the process of revision.

In the current Oslo Manual revision process, the demarcation among the different innovative activities implemented by businesses (e.g., R&D vs. patents' license acquisition or investments on new machinery) remains a crucial issue. The heart of the matter of the Oslo Manual are definitions of various types of innovations and related innovation activities. Broadly speaking, defining those (new) types and activities is the method.

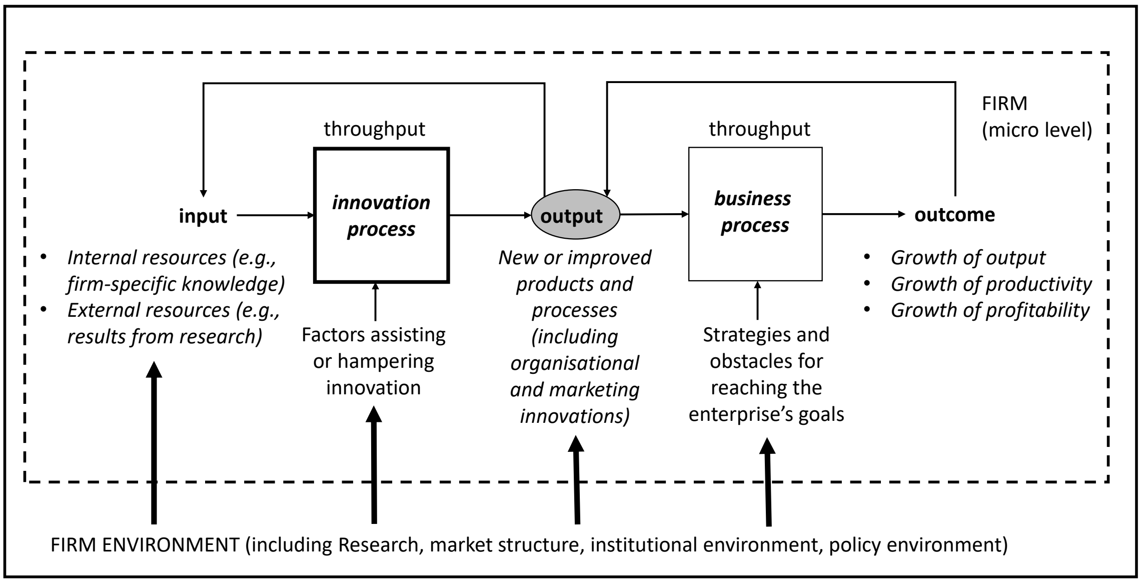
The core definition of the third edition of the Oslo Manual is that

> *"[Innovation activities] are all scientific, technological, organisational, financial and commercial steps, including investment in new knowledge, which actually, or are intended to, lead to the implementation of new or significantly improved products, processes, marketing methods or significant organisational changes. Some may be innovative in their own right, others are not novel activities but are necessary for the implementation of innovations. Innovation activities also include basic research that (by definition) is not directly related to the development of a specific innovation." (OECD, 2005).*

Within the Oslo definition, innovation can only happen as a result of a deliberate act of implementation. Such implementation can refer either to the activities needed to bring product or process on a market or to a planned effort for improving internal production process or the whole organization of the innovating enterprise. The innovation implementation is the evidence of an innovation action (which we are considering, in this paper, as the output of the innovation process). Thus, in contrast to R&D statistics, where the focus is on the input of the process (R&D expenditure, human resources) in the Oslo Manual the focus is definitely on the output of the innovation process, i.e. the innovation itself (when fully implemented).

With regard to the innovation drivers, it is quite obvious that firms are not trying to innovate per se, but rather to achieve some longer term objectives: to survive, to keep itself competitive, to be profitable, to increase sales, etc.. . The pivotal importance of innovation is based on the assumption that it is central to the growth of output and productivity (OECD, 2005) which can be seen as a firm's primary objective. Firms – as for profit institutions - only invest in innovation because they assume that it boosts (or at least preserves) their long run profitability. What matters to the scoping of the Oslo Manual is that – in the current framework - the relationship between investments in innovation (**input**) and profitability (**outcome**) is an indirect one. In an innovation process, inputs are first being transformed into new or improved products and services, processes or organisational and marketing innovations. This is the innovation **output**. It is assumed that only as a further step to the innovation implementation, these outputs will be transformed into inputs of the standard business process that uses the new processes/practices to increase productivity and the new products to increase sales and profitability. This is the **outcome**. Innovation requires both development of firm resources required to innovate, and the ability to profit from those innovations.

Figure 1. A basic conceptual model to describe the relationships between innovation input, output, and outcome


FIRM
(micro level)
throughput
throughput
input
innovation process
output
business process
outcome
Internal resources (e.g., firm-specific knowledge)
External resources (e.g., results from research)
Factors assisting or hampering innovation
New or improved products and processes (including organisational and marketing innovations)
Strategies and obstacles for reaching the enterprise's goals
Growth of output
Growth of productivity
Growth of profitability
FIRM ENVIRONMENT (including Research, market structure, institutional environment, policy environment)

Source: Dialogic (2016)

From the onset a rather pragmatic approach has been adopted by the Oslo Manual to set the guidelines for data collection. Only those innovation-related topics that were considered to be measurable were recommended for data collection. In recent NESTI meetings it has been mentioned that the focus should still be on a few key issues:

- the profiling of the key actors of innovation;
- the identification of sectors of innovation performance (preferably in line with the Frascati Manual's sector classification); and
- the assessment of various dimensions of innovation measurements (economic – or wider – dimensions) (OECD, 2015b).

In a strict interpretation this would mean to be restrictive to those areas where the innovation measurement community is able to design measurement frameworks to a full degree of detail.

It must be clearly noted that **economic results and effects (the 'outcome' of innovation) are *not* included in the scope of the Oslo Manual**.[1] Thus, the current scope of the Oslo Manual is limited to the inner left side of Figure 1Figure 1. It is the quality of the innovation process that determines how and to what extent inputs to the innovation process (e.g., expenditure, human capital, firm-specific knowledge) are being transformed into innovation outputs. The quality of the innovation process is influenced by several internal (e.g., quality of innovation management) and external factors (e.g., access to sources of information) that assist or hamper innovation.

The innovation process itself is being strongly affected by the dynamics and outcome of the business process (the right-hand of Figure 1). The propensity to innovate is linked to the 'propensity to make money': the basic assumption is that firms who are more innovative perform relatively well. Innovation indicators should therefore not only cover the propensity of firms to engage in innovation activities (and, at the meso level, to interact with other actors in the system in various ways) but also to achieve (or fail to achieve) innovation **outcomes** of various sorts – such as improved business performance.

[1] In the Oslo Manual 2005, chapter 7 is actually dealing with 'Objectives, Obstacles and Outcomes of Innovation'. Indeed, at page 20 the issue of measuring innovation outcome is introduced as follows:

> "*The outcomes of product innovations can be measured by the percentage of sales derived from new or improved products (see Chapter 7). Similar approaches can be used to measure the outcomes of other types of innovations. Additional indicators of the outcomes of innovation can be obtained through qualitative questions on the effects of innovations.*"

The point is made clear in the above-mentioned chapter 7 (page 107) where is said that:

> "*Enterprises may or may not succeed in achieving their objectives by implementing innovations, or innovations may have other or additional effects than those that initially motivated their implementation. While objectives concern enterprises' motives for innovating, effects concern the actual observed outcomes of innovations*."

It is then made clear the distinction between effects and outcomes, where effects refer to those innovation outcomes which can be observed in a survey with reference to the innovations consistently reported, that is during the same time frame (reference period). This could allow for describing effects like 'short-term outcomes' in contrast with 'long-term outcomes' which are discussed in this paper.

1.2 Where to measure outcome

Outcome can be included in innovation measurements either by (1) adding outcome variables to existing innovation surveys, or by (2) using outcome data that originates from other sources. The first option – to include some items on outcome in innovation surveys – seems to be the most obvious and accessible one. Some well-established innovation surveys such as the Community Innovation Survey and the Mannheim Innovation Panel (MIP) have indeed produced some *basic* indicators of firm performance (Eurostat, 2012) (Rammer, et al., 2016).

One advantage of this approach is that innovation input and output are directly linked to outcome, for it is one and the same firm that fills in the survey. At the same time, this is also an important weakness. The coverage of two distinctively different topics (namely innovation and overall firm performance) in **one unified survey** might be problematic. This is because the second topic will always be framed in terms of the first topic. The resulting bias might vary but could be substantial. For instance, Statistics Norway already found clear and significant differences in the results between carrying out the CIS separately and integrating it with the seemingly related business enterprise R&D survey (Wilhelmsen, Assessing a combined survey strategy and the impact of response rate on the measurement of innovation activity in Norway, 2014).

One of the practical issues might be that each topic usually requires a specialist to be correctly interpreted and statistical data carefully sourced. In the case of the CIS the main topic is obviously innovation (activities, drivers, hampering factors, etc.) and the surveys are usually addressed to the CTO and/or the manager of the R&D lab (at least in the case of larger firms), not the CFO or the CEO. This might increase the 'conceptual burden' to the respondent significantly. And, at the same time, it does not allow for reporting in detail about the overall firm performance. In the case of CIS, outcome indicators are therefore limited to two: total turnover and average number of employees, both at time t_0 and $_{t-2}$. If more sophisticated firm performance indicators are being asked from one respondent (such as the rather technical firm indicators profitability and value added in the Mannheim Innovation Panel) there is a fair chance that respondents no longer provide answers or provide incorrect estimates (Wilhelmsen, 2015).

Whether and how a potential extension of innovation surveys will influence the data provision can only be tested in practise. The common way of implementing such (cognitive) pre-tests is in a laboratory setting. However, as the results from Wilhemsen show, the tests should also be done in a field setting because the actual implementation of a survey has a highly significant effect on the answers that will be given on the survey (e.g., which respondent should be addressed, in what manner, etcetera) (Wilhelmsen, 2014) (Wilhelmsen, 2015).

The second option is to retrieve the outcome data from other, external, sources and to re-use that data for innovation measurements. This external data should then be linked in some way or another to the original innovation data sets. Although this is a more complicated option, the re-use of existing data sources has several advantages. First, it is often **more efficient** than collecting primary (survey) data. Second, it **lowers the administrative burden** on respondents (Laux & Radermacher, 2009). Third, as there is a great number of external data sources available, innovation data can be linked to a **wide variety** of outcome variables (thus not just limited to economic data). Fourth, the **data quality** (in terms of accuracy, completeness and actuality) from measurements or registrations that are dedicated to the outcome topic at hand is usually high – especially in the case of administrative data (see hereafter, paragraph 2.4).

However, as been noted, the use of secondary data is more laborious than the use of primary ones. The expected efficiency gains only occur, for instance, after major investments in the statistical architecture of the statistical agency interested to produce outcome indicators. Such investments include the implementation of **integrated statistical production systems** and the use of **unique identifiers** for enterprises. Similarly, linking confidential data with external sources requires firm measures with regard to **privacy** (anonymization, privacy by design) and **security** (authentication, access). More fundamental challenges are that the data originally has been collected for other purposes and, related to this, that the statistical agency (most often a National Statistical Institute, NSO) has less control over the data collection (Kloek & Vâju, 2013) (Laux & Radermacher, 2009).

Although these are all considerable challenges they are mainly technical in nature and in principle surmountable. We think that the costs/benefits of the second option outweigh the costs/benefits of the first option. The second option is therefore the main topic of this paper. In the subsequent chapters we will describe in more detail the practicalities of the linkage of data from innovation surveys with data from external sources.

2 Obtaining data from external sources

2.1 Availability of external data sources

There is a whole range of alternative sources of data that a NSO can use next to the traditional primary data collected in sample surveys such as, but not limited to, innovation surveys. The presumption is that ample data is **available**. There *is* a vast supply of alternative data sources but availability can be more problematic than is often assumed.

This is because availability has several layers, and each of these layers can interfere with the eventual use of the data for official statistical purposes.

First, *suitable* data should exist in the first place. Especially in the realm of business data there is a lot of data available from commercial sources but the difference in nature and/or difference in quality standards could render the data unfit for use in official statistics. Also relevant is the stability of the data holder. Private sector data suppliers might change their product portfolio, might be taken over or cease to exist altogether.

The way the data has been collected and processed also affects the methods that could be used for the linkage of the data. In the case of record linkage, individual records have to be discernible. In the case of statistical matching, characteristics of the sample have to be known (see hereafter, paragraph 3.2).

Secondly, data should also be **accessible**. That is, existing data always has to be *made* available by the data holder. The use of data might be bound to various legal conditions. In most countries, for instance, access to personal data is severely restricted. Data holders might also charge (sometimes hefty) prices for the use of their data and/or their database. In the latter case, even if the third party involved is not the owner of the data at least in Europe the data*base* might still be protected by sui generis rights. Although under the Directive on the the re-use of *public sector* information [Directive 2013/37/EU] data is in principle free of charge (or at least limited to the marginal costs of the individual request), sometimes private enterprises might have been involved in the (co-)generation, processing or distribution of the data. In these hybrid cases, commercial interests might still be at stake.[2]

Even if the use of data is allowed other conditions might apply to the ***re*-use** of the data. With regard to privacy, in many countries only data that do not allow the identification of individuals (or firms) can be made publicly available. Privacy concerns also arise when data matching is being conducted across databases that are held by different organizations, and when the matching requires identifying data to be shared and exchanged between organizations (see paragraph 3.3.1). Different price regimes might also apply to the *use* (for internal use only) and *re-use* (for wider distribution). The latter might be forbidden altogether by commercial data suppliers. Conditions for (re)use are sometimes also not particularly **transparent** and/or highly variable. A related issue is that the suppliers of the data that set the conditions should also be **accountable** for adhering to these conditions.

Finally, if suitable data exists, is accessible, and can be re-used it still has to be **comprehensible** to enable practical use. In essence, this means that the data should be machine-readable and preferably structured, that is, reside in a (traditional row-column) database. Moreover, the data should preferably be accompanied by sufficient background information (meta-data). However with the recent rise of data processing techniques that can automatically create machine-processable structures and tags the lack of ex ante structure and meta data is less of a problem. Nevertheless, the preparation of unstructured

[2] https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information

data might require substantial additional investments in hardware, software and human skills.

2.2 Types of external data

The wide variety of data than can be retrieved from external sources can be classified along various lines. One division is based on the type of data that is being stored. Buelens et al. distinguish primary, secondary and tertiary data sources (Buelens, Boonstra, Brakel, & Daas, 2012).

Box 1: Classification of data types

The classification from Buelens et al. refers to the *type* of data. However the conventional classification into primary and secundary data refers to the *method of data collection*. In the online glossary from Eurostat, the following definitions are being used:

Primary data constitute the most important inputs from among the plenitude of institutional, administrative, sample survey and/or census based information used in compiling statistical aggregates. Primary data is data observed or collected directly from first-hand experience. Published data and the data collected in the past or other parties is called **secondary data**.

Source: http://ec.europa.eu/eurostat/statistics-explained

In the classification from Buelens et al., **primary data** refers to data that is collected through a survey by an NSO.

Secondary data files are similar in structure but are not result of a sample survey. They are rather typically collected in support of some administrative process. Contrary to the survey data from primary data, secondary data (such as population of firm registers) often covers the entire population, not just a sample. Although such registers might also contain some measurement errors (e.g., due to administrative errors) the data quality is generally higher than from survey data, which is inherently prone to subjective biases from the respondents.

The notion of administrative data is generally associated with public sector information and especially with registers that are maintained by government organisations. The prevalence for the use of public sector data sources over private sector sources is understandable from a pragmatic point of view. The conditions that are attached to private sector data are sometimes quite challenging, especially in terms of accessibility. However, this situation usually refers to commercial *aggregators* of business data, not by the enterprises themselves from which the data originates (hence data that is indirectly collected, see hereafter).

Enterprises also collect vast amounts of administrative data. For instance, in most countries all enterprises are obliged to declare VAT on a quarterly basis, and corporate income tax on an annual basis (and usually in a standard format like XBRL). This also opens up possibilities for automating the data collection. Accounting software (often part of an overall enterprise resource planning suite) seems to be a particularly promising starting point in this respect (OECD, 2016).

An obvious point of action is the XBRL data exchange standard (Kloek & Vâju, 2013). In various countries around the world, enterprises are obliged to use XBRL to report to the tax authority, chamber of commerce or to the national statistical office.[3] XBRL has a taxonomy (the Global Ledger) that provides a standardized format for representing the data fields found in accounting and operation systems. The taxonomy is a modular set that can be extended with models that are geared towards special domains.[4] In principle, a specific model for innovation (most likely to be combined with research and development) could be added to the XBRL framework. This would enable the measurement of innovation activities at the level of individual projects – and thus make it possible to open up the black box of the innovation process. However, a precondition is that detailed and unambiguous definitions of innovation activities and projects are available (see Working Paper 1).

Tertiary data are usually generated as a by-process of some process unrelated to statistics or administration. Contrary to primary and secondary tertiary data usually refer to one-off occurrences (‘*events*’) rather than records that directly correspond to unit (firms, households) in a target population (Buelens, Boonstra, Brakel, & Daas, 2012). A typical example are sensor data that is being generated by devised that are connected to the internet (‘Internet of Things’), such as location data from mobile phones or logistics data from RFID scanners.

These events are only relevant to the measurement of outcome insofar some conceptual relationship exists with the units in the target population. One example of such relationships would be ownership (e.g., a household owns a device that generates relevant tertiary data about the household). Another example would be geographical overlay, where the matching is being made on the basis of coordinates. For instance, a device (e.g., a surveillance camera) generates sensor data about a certain population of units at a certain place at a certain time (e.g., a multitude of people passing through a shopping district at t_x). Tertiary data would potentially allow the measurement of innovation processes and firm performance at a level of detail that goes way beyond the project level. However similar to the use of unstructured data the use of tertiary data might also require substantial additional investments in hardware, software and human skills. One particular issue is in data preparation. Tertiary data has to be transformed into units of interest, and then target variables have to be derived.

[3] See https://www.xbrl.org/the-standard/why/who-else-uses-xbrl/ for a recent update.

[4] The current set of SBRL (v2.1, February 2013) consists of the COR (Core), the BUS (advanced business concepts), MUC (concepts that represent multicurrency information), USK (concepts specific to the US, UK, and other Saxonic jurisdictions), TAF (concepts related to the tax audit file), and SRCD (concepts that represent explicit mappings to XBRL taxonomies for financial reporting) modules.

2.3 Different ways to obtain data from external sources

The classification in primary, secondary and tertiary data does not allow for a mutually exclusive taxonomy because three intertwined dimensions are affecting it, namely *who* collects the data, *how* the data is being collected (directly at the source, or indirectly), and *what* kind of data is being collected. In general, in literature the distinction between primary and secondary data refers to the first dimension, not to the second dimension. Secondary data refers to data that was collected by someone other than the user (Donnellan & Lucas, 2013). The data in external sources is by definition secondary data since the data is not collected by the NSO itself. The third dimension refers to the *method* that is being used. Buelens et al. seem to assume a 1:1 relationship between the entity that collects the data and the method that is being used to collect the data. Primary data is survey data that is collected by an NSO. Secondary data is administrative data that is collected by another government body.

Box 2: Data collection versus method

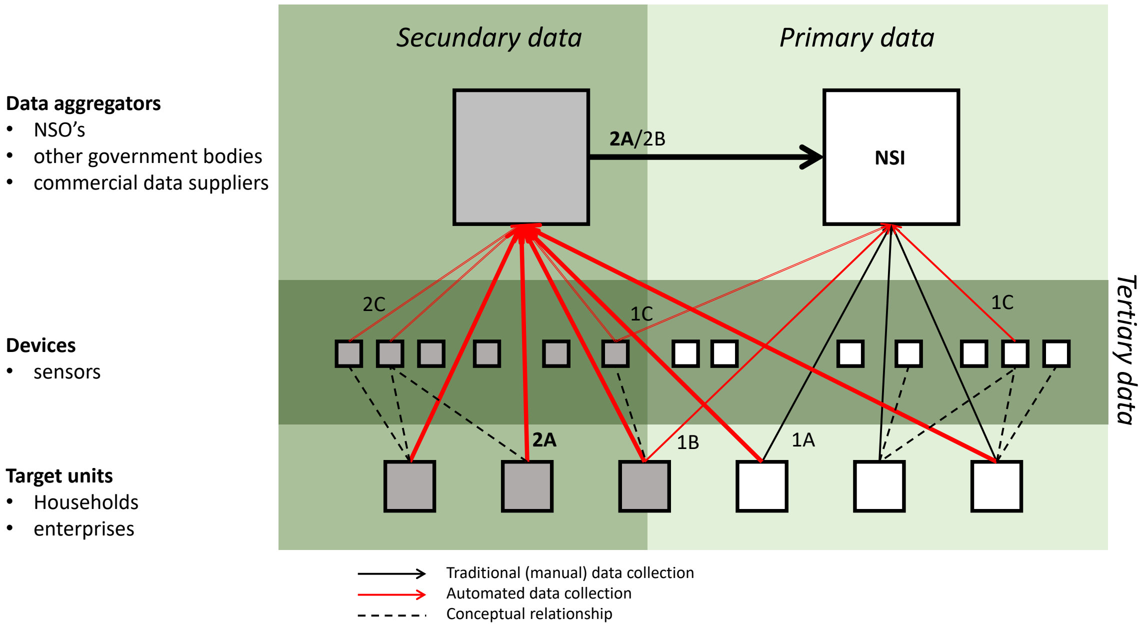
There is no direct relationship between primary and secondary data collection and method. A NSO could also collect administrative data itself. Likewise, the NSO could also use secondary data that is based on survey data. The issue with administrative data is that a NSO could re-use the data that has already been collected from enterprises by other public or private sector organisations, or to directly collect administrative data from enterprises (Boer, Arendsen, & Pieterson, 2016). The latter would refer to the aforementioned XBRL example. Thus this would be the primary data collection of secondary data.

From an efficiency point of view it would make little sense for a NSO to collect data that has already been collected by other organisations like private compilers of business registers, tax authorities and commercial data aggregators. However, the conditions that apply to the data might impede the re-use of the third party data. Re-use might be forbidden altogether (e.g., due to purpose limitation), the prices that are being charged by a commercial data aggregator might be prohibitively high and/or the conditions that apply to the re-use of the data might not be transparent. In such case the NSO might still resort to primary data collection, especially when the data collection can be (semi)automated.

Tertiary data does not refer to the first or second dimension. It is a new type of data in its own right. Although the use of the terminology seems to suggest a hierarchy – one aggregator collects data from several households or enterprises who in turn own several devices – there is only a loose relationship between the target units and the devices. Essentially, the internet of things is a layer *between* the target units and the data aggregators. A NSO could directly collect data about (a sample of) target units or indirectly via sensors that have a conceptual link with the target units. Again, a NSO could collect the tertiary data itself (primary data collection) or retrieve it from a third party (secondary data collection).

In the figure below the various options that have been described in this paragraph are visualized.

Figure 2. Various ways for an NSO to collect data



Source: Dialogic (2016)

1A refers to the notion of primary data from Buelens et all. It is a primary data collection by a NSO by means of a survey (Buelens, Boonstra, Brakel, & Daas, 2012). 2A refers to their notion of secondary data. A data aggregator (another government body, a commercial data supplier) collects administrative data from enterprises. In the particular case of public registers the entire population of enterprises is usually covered – as registration is often obliged by law. Due to the resulting large scale, the data collection is often largely automated. NSO's could obtain the administrative data from the data aggregator. If the conditions to re-use the data are unfavourable a NSO could also decide to directly tap into the administrative data from enterprises (1B). This also usually requires automated data collection. Tertiary data – sensor data – is a separate layer. A NSO could collect such sensor data itself (1C) or obtain it from a third party who has collected the data (2C). The events that are being captured by the sensors are only relevant insofar they can be conceptually linked to (a set of) target units.

2.4 Re-using administrative data

The use of tertiary data is a promising development and the exploitation of big data has become an integral part of most NSOs' strategic planning, as for the Eurostat's ESS Vision 2020 (Eurostat, 2014). Nevertheless, while awaiting for common classifications, definitions and formats, big data is difficult to harmonise and therefore at the moment still difficult to be effectively used in existing statistical structures. Its huge volume and velocity make storage challenging, while new tools and statistical methods need to be developed for its processing, analysis, anonymisation, and visualisation (Augustyniak, 2015).

In contrast, the structure of secondary data is quite similar to primary data. An additional advantage vis-à-vis primary (and tertiary) data is that secondary data (e.g., administrative registers) often cover the complete target population (Buelens, Boonstra, Brakel, & Daas, 2012).

In this chapter we will therefore focus on the linking of innovation survey data with secondary (micro) data, more precisely administrative data from public registers[5]. The linking of innovation input and output data with outcome data from third parties (most notable administrations) is most in line with developments taking place in the international statistical community (e.g., ESS) and it is in fact already widely practiced. The potential range of administrative sources that could be used for statistical purposes is large and growing (Dias, 2015). Examples of administrative sources that could be relevant to the study of the outcome of innovation processes are tax data, published business accounts, licencing systems, building permits social security data, education records.

The use of register data has a long history in statistics. In fact, administrative sources (e.g., census) have been the basis frame from the very beginning of official statistics.[6] However the use of administrative data only really took off from 1980 on. The trend is directly related to the introduction of integrated information systems within NSOs. Population frames (e.g., business registers) have already been used for decades as a sampling frame in surveys but these frames can now become the backbone of the integrated system ('information warehouses') to which all information could be somehow linked (Kloek & Vâju, 2013).[7]

However, the use of administrative data does brings a number of challenges with regard to data quality, the management of statistical production systems, and the required statistical infrastructure (Eurostat, 2003) (Laux & Radermacher, 2009).

Challenges with regard to **data quality** mainly arise from the fact that the (secondary) data have initially not been collected and processed for statistical purposes. Thus comparability issues might occur due to the fact that different concepts, methods, planning and legal regimes have been used in the collection of administrative data (Laux & Radermacher, 2009). The administrative purpose must at least be a good approximation to those required for statistical purposes (Eurostat, 2011). This means that the NSO – as a re-user of the administrative data – should at least have a good insight in the way the secondary data has originally been collected and processed. For instance, sampling and non-sampling errors should be systematically documented. Preferably, standardized meta-data should be used

[5] This is option 2A in Figure 2.

[6] In The Netherlands for instance, data for the regional income survey (established in 1946) is completely obtained from tax and social benefits registrations.

[7] A particular relevance here will have the work done to establish the EuroGroups Register, see http://ec.europa.eu/eurostat/statistics-explained/index.php/EuroGroups_register

(e.g., the *Euro SDMX Metadata Structure*[8]) that describes the integrity, quality and other aspects of the data (Laux & Radermacher, 2009).[9]

Challenges with regard to the **production system** obviously arise from the fact that the data is not produced by the NSO itself but by another administrative body.[10] This means that the NSO has less direct control over the data collection process. Addressing the aforementioned quality issues therefore requires a strengthened coordination between NSOs and the organisations that produce the data. Such coordination could for instance be formalized via protocols (Laux & Radermacher, 2009).[11] A more far reaching coordination could also involve the NSO in the design and configuration of the administrative data, as well as in the quality review of the data. One of the main obstacles for the re-use of commercial data is that NSOs have usually little or no control over the original production of such data. Since commercial aggregators have little incentive to be transparent about the quality of their data, it is often difficult to assess the quality of the production process and the resulting data.

Confidentiality of administrative data is another critical issue for the re-use of administrative data. Administrative data (e.g., person data or business data) are usually confidential. Consequently, proper legal arrangements and security measures should be made. With regard to the legal dimension, personal and business data is often protected by the basic principle of purpose limitation. This means that an exemption should be made for statistical purposes. Furthermore, the NSO should ensure that the identity of individuals or firms is protected at all times. Data protection legislation often demands additional steps: anonymisation and dealing with access right (Kloek & Vâju, 2013). Anonymisation should ensure that the data that is being published should never be deducible to a specific individual or firm. Access to secondary data is usually restricted to a known of approved users whom credentials are known (that is, they have been screened) and who (re)use the secondary data is in a controlled environment (e.g., a Research Room). The screening (and subsequent permission to access) could be done by the government agency that holds the administrative data or it could be delegated to the NSO.

[8] http://ec.europa.eu/eurostat/data/metadata/metadata-structure

[9] Such upstream investments in data quality greatly facilitate the re-use of administrative data downstream (e.g., by an NSO). Because the costs are the lowest at the source from a welfare economics point of view this is also the most optimal solution (te Velde, 2007). Alas the benefits do not directly require to the administrative body that produces the data. Which brings us to the next challenge, namely that the NSO has limited influence about the production of the administrative data.

[10] such as, for instance, the tax authority in the case of relevant secondary data on outcome, i.e. firm performance.

[11] Such protocols should at least cover the description of the administrative sources (again, preferably using standardized meta-data) and the NSO is notified about any changes to the administrative system and data collection.

3 Linking innovation data with outcome data

In the previous chapter we have dealt with the availability and collection of data from external sources, with a focus on the re-use of administrative data. In this chapter we will cover the subsequent step, namely the actual linkage of the different data sources.

3.1 Integrating other data sources in a statistical infrastructure

Data set integration

Data integration is defined broadly as the combination of data from different sources about the same or a similar individual or institutional unit. A *precondition* to data integration is the harmonization of the data *sources* or **dataset integration**. Data cleaning and standardisation are crucial preparatory steps to successful data matching. This integration involves the adjustment of data sources at various hierarchical layers.

At the physical layer at the bottom, differences in *technology* need to be overcome. These differences can exist in hardware and software (operating systems, databases' structures and formats). However by using common standards as intermediaries, no further fine-tuning at the technology layer is needed.

The exchange of data subsequently assumes to rely on a common structure. That is, the *syntax* of the records – the way in which entities are represented (e.g., in terms of formats, measurement units, ranges etcetera) need to be harmonized. If there is no common syntax the data needs to be extracted from the data source as raw data and then be cleaned, standardized and eventually parsed into predefined formats and data structures. The objective of parsing is to segment each output field into a single piece of information (e.g., `COMPANY NAME`, `LEGAL FORM`) rather than having several pieces as a single field or attribute). Standardisation is particularly problematic in the case of names. Names are often spelled differently and/or companies (especially large ones) operate under many different names and have many subsidiaries.[12]

Finally, differences in the *semantics* of the data need to be overcome. This requires the fine-tuning of meanings, concepts and definitions.[13] Coming to terms with semantics is often the

[12] Much progress has nevertheless been made in name matching. One practical example is the OpenCorporates database (https://opencorporates.com) that has nearly 100 million companies and that is run by just a couple of database administrators.
[13] At the European level, the EuroGroups Register (EGR), that is part of the FRIBS initiative, works towards uniform definitions of enterprise information at three levels:

most difficult step in the harmonization process because it requires detailed adjustments and a deep understanding of the domain from which the data originates. This involves a lot of coordination between the NSO and the organisation that produces the data. This is one of the main hindrances for the re-use of data from commercial aggregators (see before, paragraph (Sturgeon, 2014).[14]

The integration of other data sources in the statistical infrastructure of a NSO would at least require an information system that it would be able to identify and link population units (e.g., individuals and enterprises) across different internal and external data sets. This does not necessarily require a centralized register but at least a **series of compatible registers**. In the case of enterprises this requires a **unique identification numbering system** managed by business registers and used for every statistics included in micro-data linking programs (Sturgeon, 2014).

Micro integration

Micro integration is a complementary part of the overall data integration process. The purpose of micro aggregation is to compile *better* information than would be possible by using the sources individually (Bakker, 2011). Micro data could be linked (e.g., individual data could be enriched by adding additional information to units) without improving the overall data quality of the integrated data set (Al & Thijssen, 2003). Micro integration is intended to improve the outcome of record linkage and/or statistical matching. It is applied in situations where variables from different sources may have values that are incompatible or inconsistent with each other at the unit level. Some data sources have a better coverage than others or are just more reliable than others. In many cases there might even be conflicting information between sources at the record level. The basic idea of micro integration is to take the best data quality of several data sources. The final goal is to generate a dataset in which all perceived incompatibility or inconsistency had been removed (Dias, 2015).

Proper micro integration is only possible in a fully integrated statistical infrastructure. In the ideal case, for a limited number of basic units (e.g., individuals, businesses, buildings) statistics are being complied by matching, editing, imputing and weighting data from the combined set of administrative registers and sample surveys. Since there are inevitably differences between data sources, a micro integration process is needed to check the quality of the data and to adjust for incorrect data. Ideally, the data source with the highest quality for a particular basic unit or variable is used as the overall quality benchmark for the

1. **enterprise groups**: identity, demographic characteristics, the structure of the group, the group head, the country of global decision centre, activity code (NACE), consolidated employment and turnover of the group.
2. **enterprises**: identity and demographic characteristics, activity code (NACE), number of persons employed, turnover, institutional sector;
3. **legal units**: identity, demographic, control and ownership characteristics.

entire system. Hence an important task of a NSO is not only to identify the widest range of available data sources (see chapter 2) but also to assess the strengths and weaknesses of each particular data source that could potentially be added to the system. Eventually, then, micro aggregation could provide far more reliable results because the integrated data are based on an optimal amount of information (Dias, 2015). The coverage of (sub)populations is also better because missing data in one data source can be filled by data from other sources (e.g., by statistical matching). Finally, the consolidation of data sources makes sure that a uniform figure is published for a specific unit or variable.

3.2 Linking micro data

Overview of linking methods

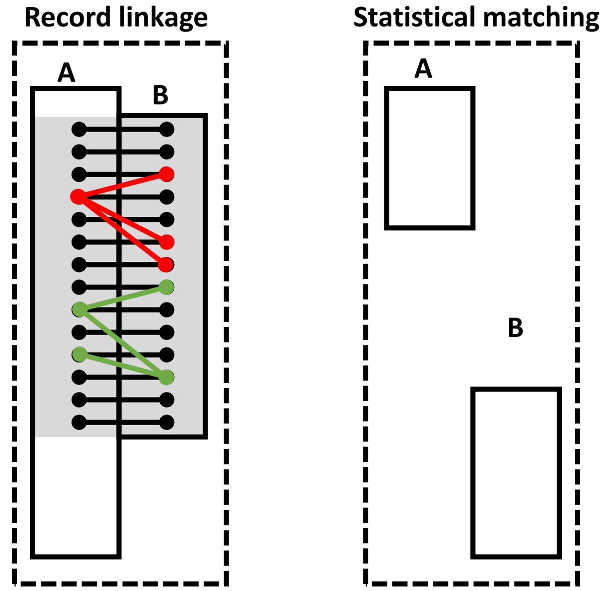
There are two basic methods to link data:

(1) **Record linkage**: a different set of information on the same unit (hence *identical* records);
(2) **Statistical matching**: information on a unit with the same characteristics (hence a *similar* unit).

Method (2) can be applied at either the macro level (of groups) or the micro level (of individual units), method (1) is by definition solely applied at the micro level. In this chapter we will focus on data linkage at the micro level.

The two methods are applied to different types of input data. Record linkage is used when the data sets that are linked have at least a partial overlap in units. In the particular case of register data, the data set usually covers the entire population. Thus there is a near complete overlap between the two data sets. The units are *directly* linked at the record level. Records can be linked one to one, one to many or many to many (respectively the black, red and green ties in Figure 3). Statistical matching is used when there is no overlap between the units (hence there are two independent samples). In this case, statistical matching is the only possible method to link units.

Figure 3. Various ways for an NSO to collect data



Record linkage

Same population, overlapping groups of units

Statistical matching

Same population, two independent samples, no overlapping units

Source: (Dias, 2015)

Record linkage and statistical matching also generate very different types of output data. The units in the output data from record linkage refer to real-world entities, that is, individuals or firms that really exist. Obviously, this has severe privacy consequences (see paragraph 3.3). Statistical matching at the micro level combines two different real-world entities into a new *virtual* unit (hence the alternative label *synthetic* matching). This means that there are less privacy concerns. The drawback of statistical matching is that the combination of the data is tailor-made for every analysis. This means that the virtual units that are being generated in the output cannot be re-used for a new analysis – unless the same common variables are being used. Thus record matching is a more flexible and permanent solution for data integration.

There is a third method that uses the characteristics of both methods, namely

(3) **Micro aggregation**: a grouping of (a small number of) records based on a proximity measure of variables of interest.

The purpose of micro aggregation is not to link data but to *preserve the privacy* of the individual real-world entities (hence we will further describe this method in the paragraph on privacy and security, paragraph 3.3). Privacy is preserved by releasing the *aggregates* as output units instead of individual record values. Since these aggregates are also synthetic and the grouping is based on the specific variables of interest, micro aggregation has the same drawback as statistical matching. The virtual units cannot be re-used for a new analysis unless the same variables are used for the grouping.

Record linkage

Record linkage is the task of identifying and matching individual records from different databases that refer to the same real-world entities or objects. Records are matched on the basis of a unique unit identifier.

In the ideal case, there is already a generic **unique ID** available (e.g., PIN, business registration number, VAT number) – and it is allowed to use the number as a key to couple data. In many countries this is not the case but there are workaround with anonymous keys (see paragraph 3.3.1).

If such a unique key does not exist, records can still be matched by combining several fields of the record, a.k.a. characteristics of the unit.[15] It is the specific *combination* of the supporting fields ('partial identifiers') that needs to be unique, not the fields itself. These partial identifiers should preferably be unique for each unit, available for all records (universal), stable (permanent), recorded easily and without errors (accurate and non-sensible), and simply verifiable (transparent) (Dias, 2015). This obviously assumes a high data quality.

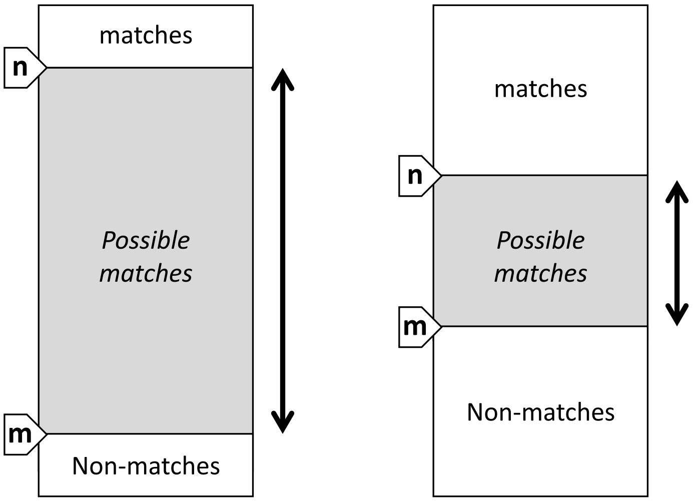
However, in the frequent presence of administrative errors (e.g. due to difficulties with the standardization of names or due to highly dynamic population thus outdatedness of records) it might not be possible to use **deterministic matching**. If data is noisy and contains random errors but there is an array of partial identifiers that could be used for blocking and record matching, one can still resort to **probabilistic matching** (Fellegi & Sunter, 1969). In the latter case, matches (A=A) or non-matches (A≠B) between individual records are not perfect but instead have a probability ('similarity value') between 1 (A=A) and 0 (A≠B). For each candidate record pair *several* attributes are generally compared, resulting in a 'comparison vector' of numerical similarity values for each pair.[16] If the probability is close to 1 there is a *possible* match between the two records (A=a?).

In probabilistic matching, the critical problem is to determine the optimal threshold scores for matches (m) and non-matches (n). Obviously, the higher m and the lower m, the larger the number of *possible* matches that need to be further scrutinized. The threshold scores can either be determined by trial and error or by using a model based approach (as proposed by Fellegi and Sunter). The optimal setting depends on the specific characteristics of the data sets hence needs to be tailor-made. The determination of the threshold scores is an iterative process in which the number of type I (false positive, A=B) and type II (false negatives, A≠A) errors are minimized.

[15] In a similar vein, by combining several fields the identity of individuals or individual firms could still be deduced from anonymised data. When multiple high dimensional datasets are being coupled the intersection becomes so small that anynomity can no longer be guaranteed (k-anonymity) (Montoye, Hidalgo, Verleysen, & Blondel, 2013).

[16] Compare the use of propensity scores in statistical matching, see hereafter, paragraph 3.2.3 and footnote 22.

Figure 4. Range of possible matches with strict and lenient threshold values for matching



To assess the accuracy of the matching, in information retrieval the measures of *precision* and *recall* are often used.[17] Recall is the proportion of positive cases that were correctly identified (or pairs correctly matched, hence recall refers to pairs completeness).[18] Precision (or pairs quality) is the proportion of the predicted positive cases that were correct. Both recall and precision should be high but there is a trade-off between the two measures.

If data sets are very large the complexity of the data matching process (and hence the computational capacity needed) can be reduced though the use of data structures that facilitate the generation of candidate record pairs that are *likely* to correspond to matches. This is being done by dividing the two data sets into comparable subsets of units ('blocking'). A disadvantage of blocking is that type II errors might occur: matching record pairs are classified as non-matches because they are not members of the same block (Dias, 2015). In general, there is a trade-off between ex-ante data structuring (for instance, by blocking) and computation cost. With the increasing availability of computational power (and cost-efficient solutions such as cloud computing) pre-structuring of the data is no longer needed and flexibility in the data can be maintained.

Whereas automatic matching is used by many NSOs as a cost-efficient approach for record matching in bulk, clerical intervention is still needed for the proper resolution of controversial matches. The range of automatic detection could be increased – and thus the deployment of expensive human agents minimized – by optimizing the data preparation (e.g., by additional investments in the data wrangling process) and especially by improving the intelligence of the automated agents. However the latter also comes at a high price: the development costs of the automated system will rise sharply (Brennenraedts & te Velde, 2012).

[17] The full set of measures and derived measures for assessing the accuracy of information retrieval or matching is given by the confusion matrix or error matrix (Provost & Kohavi, 1998).

[18] Recall (or true positive rate, *TPP*) = matches / (matches + false negatives). Precision (*P*) = matches / (matches + false positives).

The drawback of linking records without unique ID is it that it has to be tailor-made every time two or more data sets are being coupled. Moreover, record matching becomes more difficult when the number of data sets increases. The most efficient solution is therefore to assign a unique key to a record once a definitive match has been made.[19]

When a system of registers with unique identifiers has been established, a NSO could in principle combine any register (and census) at any time. This does however require a careful management of the IDs within the statistical infrastructure. For instance, for each of the base registers a **standardized population** or population frame has to be created (Wallgren & Wallgren, 2011).[20] Thus, changes in external registers which could affect the matching precision should be closely monitored. Old and new IDs should for instance be included in a cross reference table together with the reference time when the change occurred.

In the specific case of firms the dynamics in the population are high, thus frequent updates are needed. At the same time, administrative sources for business statistics (such as Chambers of Commerce directories) usually are mainly interested in registering the changes of status and/or in reporting the formal (legal) status of a firm, mostly based on ownership, rather than monitoring its evolution overtime in terms of size and economic activity. For statistical purposes the economic continuity is more important (Kloek & Vâju, 2013). This means that different **continuation rules** will have to be adopted for the base register and the standardized population of firms that is being used as the backbone of an integrated information system of a NSO.[21] The standardized population is therefore an accurate but not an exact copy of the business register.

Statistical matching

Statistical matching techniques are used to combine information that is available in distinct data sources (e.g., a CIS dataset and an external data source with outcome data) that refer to the same target population (i.e. firms or a specific subset of firms). An important condition is that the units in the two data sets *do not overlap*, hence direct matching via record linkage is not possible. Note that in the case of public registers the coverage of the population is usually nearly complete, and often unique keys are available as well. Consequently for matching a set of innovation micro data with register data, record linkage is usually the most appropriate method.

Thus, if there is no overlap between the data sets that need to be combined, record linkage cannot be applied but the critical condition for statistical matching ('no overlap') is exactly

[19] This operation can also be applied within a single database. The purpose of record matching – and the subsequent assignment of one unique key to both records – is then to detect and remove duplicates.

[20] In the case of Statistics Sweden, special units within the NSO are responsible for one of the base registers (Wallgren & Wallgren, 2011). In other cases, other administrative authorities or dedicated public agencies might be responsible for the administration and maintenance of one or more base registers.

[21] For instance, firms that are still registered as active firms in the business register but that not have submitted VAT declarations for, say, the last two quarters could be dropped from the population frame of firms that is being used by the NSO.

met. The purpose of statistical matching is to study a relationship among *variables* that only occur in either one of the data sets (Y in data set A and Z in data set B). The actual matching is being done on the basis of a variable that occurs in both datasets (the joint variable X). The first condition stipulates that Y and Z are conditional independent given Z. That is, Y and Z should *not* be jointly observed.

In the *micro* approach of statistical matching, a completely new micro-data file is created where data on all the variables is available *for every unit*. Based on the common variable X, for every unit with variable X variable Z is imputed or, the other way around, Y is imputed into units with variable Z as well.

The joint variable X is used to select samples of the two data sets that have the greatest resemblance, that is, as much as possible similar covariate distributions. The strategy is to select those samples for the bias to the covariates is minimized (Stuart, 2010).[22] The matching can be done multiple times and the matched samples with the best balance – that is, the most similar samples of data set A and data set B – are chosen as the final matched samples.

The crucial point is that the optimal composition of subsets of units from A and B is dependent on the target variables Z and Y (that are being studied). The choice of the target variables influences all subsequent steps in the matching process (Dias, 2015). This means that the matching should be tailor-made for each specific analysis and furthermore, that the variables Z and Y should be a priori known. This makes statistical matching much less flexible than record linkage. In the latter case, the selection of the subsets can be optimized for the specific purpose of the study and new variables can always be added later on.[23] Data that is being matched on the basis of individual records allows the reuse of existing data sources for new studies. This is usually not to case for data that is combined on the basis of statistical matching.

[22] For the calculation of X, propensity scores are often the best available method (Rosenbaum & Rubin, 1983). Propensity scores summarize all of the covariates into one scalar: the probability of being treated. The propensity score is defined as the estimated conditional probability of a unit to belong to data set A (dummy value = 1) or data set B (dummy value = 0). A logit or probit model is estimated with the dummy as dependent value, and the common variables X as independent variable, obviously including the regression constant. Then, for each recipient record a donor unit is searched with the same or the nearest estimated propensity score. Next to mixed methods such as propensity scoring, other types of micro-matching methods are hot deck imputations or regression based models. For an overview, see (Eurostat, 2013).

[23] The experiences with opening op public sector information have actually shown that it is very difficult beforehand to envisage how data will be used. In general, there are many more creative combinations possible than is initially thought.

3.3 Privacy and security aspects of data integration

Record linking

In the IT tradition flexibility is the key. In theory, information systems are designed in such a way that they are robust to future developments. Record linkage is the most flexible solution. The golden standard is to link at the level of individual records that have a universal definition and a globally unique ID (e.g., UUID's in software).[24] In principle, any new data source or new attribute can - ex post - be linked to the set of units.

The possibility to link across any data set is also the biggest drawback: this comes with major security and privacy concerns. Linking confidential data at individual level across data sources and data holders is a hazardous venture in terms of privacy and security. This is the price one has to pay for maintaining flexibility.

There are several measures to protect privacy of personal and confidential data but privacy can never be completely guaranteed. At some place in the data infrastructure a coupling with the original unencoded and unencrypted records must be made, thus there is always a theoretical possibility to trace back information to a specific individual or a specific firm.[25] The only fool proof solution would be to use synthetic units. The integrated data would no longer be suitable for *administrative* purposes (as it does not refer to real-world entities) but it could still be used for policy making and research purposes, which makes up about 50% of all usage of business statistics (Eurostat, 2015). However one can never be sure to what extent the statistical matching or micro aggregation influences the eventual results of the analyses of the linked datasets. The biggest drawback is the loss of flexibility. The only basis of an integrated statistical infrastructure is record linking.

The two basic measures to protect privacy are to *limit access* to the data or to *conceal* the data. In many countries, only data that do not allow for the identification of individuals (or individual firms) can be made publicly available. At the input side this means that re-use of data sources with personal of confidential data is forbidden altogether (hence the data source is not available) unless an exemption has been made for re-use for official statistical purposes. At the output side this means that data should be aggregated in such a way that it is impossible to reverse engineer the aggregation process (Torra & Navarro-Arribas, 2015).

Ideally, control of the data is at the lowest level. Thus, it is the original data holder (usually the producer or the owner) that keep a full control of what data to release, to whom and in what manner. Thus eventual anonymization of the data should preferably already been

[24] See http://www.ietf.org/rfc/rfc4122.txt

[25] The NSO typically has a master table in which the original identifiers (company code) is being linked to an anonymous ID that enables the linking of records without revealing the identity of the firm. The actual linking can also be done without revealing the identity of the firm, for instance by using hash tables. Still, even if anonymous IDs are being used users would in principle be able to deduce the identify of individual firms by combining various fields. When multiple high dimensional datasets are being coupled the intersection becomes so small that anynomity can no longer be guaranteed (Torra & Navarro-Arribas, 2015).

done by the data holder, prior to the data exchange with the NSO. However identifiers still have to be known to the NSO otherwise records cannot be linked. In case unique IDs are not available the matching of databases needs to be done on the basis of partial identifiers (see paragraph 0). However the most suitable (universal, stable, accurate) auxiliary fields to be used for matching usually also contain most personal data (names, addresses, dates of birth or establishment etc.). If there are unique IDs available (social security numbers, business register numbers) in many countries it is not allowed to use these keys to link with external data sources – ironically because they make linkage so easy. There is a workaround for NSOs, namely to use a parallel set of unique IDs. In this way, registers can be linked internally, within a NSO, without using the original ID as a key. Thus, in the linking of data the identity of the firm does not have to be revealed. However, at some place in the NSO a linkage table with the original IDs still needs to be kept. Security is maintained by physically limiting access to the data. That is, only a selected number of individuals are authorised to get access. In a similar vein, a NSO could limit the access to the anonymised micro data to on-site use only and require prior screening from the external users (see next paragraph, 0).

The best option to preserve the confidentiality of the micro data would be to use **privacy by design**. Again, ideal control of data should be kept at the lowest level. Currently, this is the level of original data holders. But these are still data aggregators. The actual lowest level is that of individuals or firms, i.e. the real-world entities to which the records in the micro data refer (the target units in Figure 2). There are a number of recent developments that enable to transfer micro-data control to the lowest level of individual users. We are referring here to the emergence of **distributed information systems**, with distributed hash tables and the block chain technology as the most notable implementations. These technologies use the network *as a whole* to verify the legitimacy of data operations rather than some kind of centrally-authorized actor (such as a NSO or a trusted third party, TTP).

The use of distributed ledgers allows citizens and firms to manage the access to their data and to know who has actually accessed the data. The security and accuracy of the ledger is maintained cryptographically (using proof-of-work or proof-of-stake schemes) to enforce strict access control. In essence this allows for the consensual use of personal or confidential micro data in anonymous form for collective intelligence, such as re-use of statistical data for research purposes. Blockchain applications such as Ethereum, for instance, enable to use of so-called 'smart contracts' to create a permanent, publicm transparent ledget system for compiling all sort of personal data (e.g., rights data, digital use data etc.) (Buterin, 2014).

In essence, this allows for the consensual use of personal or confidential micro data in anonymous form for collective intelligence, such as re-use of statistical data for research purposes.

Given the flexibility of record linkage this method is more robust to future developments such as the rise of distributed information systems than statistical methods like micro aggregation.

Implementing data linking in practise

Provided that the statistical infrastructure of a NSO would meet the basic requirements and that a centralized company register with unique IDs would exist, there seem to be no major hurdles to link input data (e.g., CIS data) with outcome data (e.g., on firm performance) from third parties (usually other administrative bodies such as Tax Authorities). In fact, many NSOs already facilitate such linking of data.

The key question for the implementation of the data linking will then be who will actually perform the linkage, and how the confidentiality of the micro data can be safeguarded.

Limiting access to micro data is one measure. It means that NSOs have to screen every potential user of its micro data. NSOs then have to make sure that no confidential data is being brought outside the (physical and/or virtual) 'Research Room'. An alternative would be if the NSO provides the linkages and only publishes the aggregated data – this is basically what is happening if micro aggregation is being used (see next paragraph, 0). However it is often a more practical approach when the end user itself (e.g., a researcher) links the data sets rather than the NSO.

Based on a broad stocktaking of current practices we found four basic implementation schemes:

1. **The NSO makes the linkage in-house and publishes the linked data**. We found only one case in our sample, namely New Zealand. In the annual Business Operations Survey Statistics New Zealand publishes cross tables on types of innovators (CIS alike) x business performance (income, expenditure, profit). The linkage is probably based on micro aggregated data, not an a direct linkage between individual records (see hereafter, paragraph 0).
2. **The NSO makes the linkage in-house but leaves it to third parties (e.g., external researchers) to publish on the data**. This is also a rare scheme. Israel has a linked set with innovation and value added data. The data set is available for internal use or for the Research Room from the Central Bureau of Statistics.
3. **The NSO provides data sets that can be linked but leaves the actual linking (and publishing) to third parties.** This is the most common scheme. At least the NSOs of Australia, Norway, Sweden and The Netherlands, provide such data sets to external researchers (usually via their Research Rooms). Since firms have a unique anonymous ID across all data sets linking the data is not a major challenge and is, in fact, being done on a frequent albeit ad hoc basis by academic researchers and research consultants who have been granted access to the Research Room.
4. **The NSO does not provide data sets that can be linked.** For a number of practical reasons, there are quite some NSO's who do not provide data sets that can be readily linked. First, relevant data sets might not be available at all. Secondly, linking of data across authorities might not be allowed due to privacy regulation (e.g., Sweden).

> Thirdly, the country does not have a centralized company register (e.g., Germany) and/or unique IDs are missing (e.g., Belgium, Germany).[26]

The main reasons why type 3 is that most widely found are three: (1) assuming an unique ID exists linking data sets is a trivial operation. It has little added value to offer this service; (2) a NSO has many different data sets hence many possible linkages are possible; (3) often a researcher has specific research questions thus needs a specific selection of variables, industry sectors etcetera. In short, although a NSO could very well offer one big generic linked data set (of which every researcher can make its own cross-section) this is often not the most practical solution.

The prevalence of type 3 is invigorated by the way most NSOs are currently funded. Linking input with output data is usually not part of the mandate of a NSO and it is therefore not financed from its institutional funding but rather on a project-funding basis. On the other hand, the linkage is usually made upon request from third parties (consultants or researchers – the latter could also be internal NSO staff) and paid per request. Alas on the macro level this has the disadvantage that it preserves the existing fragmentation in innovation research. For instance, results from incidental research projects are ill-structured for cross-referencing. Moreover, in some countries external researchers noted a further increase of fragmentation due to the fact that research groups focus on maximising output on the data that they bought. Hence they might be less willing to share the data then before.

Micro aggregation

Micro aggregation is a statistical disclosure technique that has been introduced by Eurostat researchers in the early 1990 (Defays, 1997). The basic principle is to split a population in as small as possible groups so that the resulting aggregates cannot be traced back to an individual unit of the population while these composite units still behave similar to the original individual units. That is, a 'virtual person' or a 'virtual firm' is being created that has the same characteristics as the real firms of which the composite unit is constituted. This could be regarded as the statisticians' way of '**privacy by design**'. Provided that multivariate (rather than univariate) micro aggregation is being used, confidentiality of the data is ensured because it is embedded into the very design of the statistical technique.[27] That is, the anonymisation of the data is already been done *at the source*, before datasets are being merged. This is a definite advantage over the IT-based method or record linking that uses individual records that refer to real-world entities.

[26] Obviously, in the latter case matching on partial identifiers could still be used to link records. However this requires much more efforts than record linking on unique ID's.

[27] Micro aggregation ensures k-anonymity only when multivariate micro aggregation is applied processing all the variables of the data file at the same time. Otherwise, this is not ensured. In fact, it is often the case that k-anonymity is not ensured. This is so because the set of variables is often partitioned, and micro aggregation is applied independently to each partition element. This is done to achieve a lower information loss (higher data utility) than when applying it to the whole set. In this case, a trade-off has to be found between the information loss and the disclosure risk (Torra & Navarro-Arribas, 2015).

At the same time, similar to micro integration, this is also the biggest drawback of the method. Because the linkage is being done at the level of 'virtual firms', not at the level of individuals, quite a lot of flexibility is been lost. Direct matching with other years for the same datasets (e.g., to facilitate panel designs) or with new data sources (e.g., that contain other types of outcome variables) is no longer possible. The matching is always on a limited number of generic background variables (e.g., firm size, sector, year) that are defined beforehand. Provided that the size of the micro aggregated cells would be kept reasonably small, there will be a lot of cells (a.k.a. 'virtual firms') available for analysis, hence it will be possible to make a lot of different cross-sections.

Although the micro aggregation can be done on the basis of generic variables this is not the optimal solution. Ideally, the micro aggregation should be **tailor-made** for the type of analysis that will eventually be conducted on the micro aggregated data (Lamarche & Pérez-Duarte, 2015) – compare statistical matching (paragraph 3.2).

For instance, in the ESSLait project the data has been prepared for the specific purpose of conducting ICT impact analysis (Hagsten, Polder, Bartelsman, & Kotnik, 2013). The input variables have been chosen beforehand and conveniently fixed. This is done because there is a trade-off between the number of variables that are used to compute the Euclidean distance and the errors that occur due to the micro aggregation. One of the main difficulties in micro aggregation is the clustering process of how to select similar units ((i.e. reducing intra-cluster variance as much as possible) while ensuring a sufficient but not too high number of units in each cluster (Lamarche & Pérez-Duarte, 2015). An optimal clustering would require that the variables that are used for the calculations are fine tuned to the specific analyses which will be performed. However, this assumes that the purposes of the analyses are already known beforehand, and this is obviously not always the case.[28]

Micro aggregation has another application besides preserving privacy. A condition for record linking is that the definition of the units is harmonized. In practise, the condition is often not (yet) met. For instance, the definition of the basic unit ('enterprise') differs across countries because data is collected in different ways.[29] In this case, micro aggregation can be used to group units that are not exactly identical into synthetic units that are 'similar' across data sets.

[28] The experiences with opening op public sector information have actually shown that it is very difficult beforehand to envisage how data will be used. In general, there are many more creative combinations possible than is initially thought.

[29] A relevant development is the changes that have been made to the FRIBS roadmap. In response to the concerns expressed by many National Statistical Institutes, it has been decided to exclude an update of the definitions of statistical units from FRIBS. Instead and in parallel, Eurostat has launched immediate measures for helping Member States in complying better with the existing statistical units definitions in each of the statistical domains and the Business Register (Eurostat, 2015b).

4 Conclusions

Outcome in terms of improved business performance is the raison d'être for the measurement of innovation. The question how to measure innovation outcome indicators such as growth of output, productivity and profitability, should be addressed in the Oslo Manual beyond the more traditional approach focused on 'innovation effects'. The main distinction between effects and outcomes is the time reference: effects can be observed during the same reference period used to provide a time frame for innovation measurement, while outcomes can emerge after it, according to specific time-lags.

Questions of innovation effects – either actual or expected ones – are usually asked in innovation surveys. Similarly, the most obvious – and straightforward – way to measure outcomes is to include as well questions on outcome indicators in surveys.

The challenge is twofold. On the one hand, the observation period will have to be limited to a few months after the implementation of an innovation (thus excluding longer term outcomes). On the other hand, there is a high risk of confusing effects and outcomes. This is in fact the risk affecting the introduction of outcome/effects questions in some established innovation surveys such as the Community Innovation Survey and the Mannheim Innovation Panel. Moreover, this approach limits the measurement of outcome to a few basic outcome items and it does not expand the scope of the measurements. As a conclusion, it can be said that the combination of the measurement of innovation and outcome in one unified survey might introduce substantial biases.

A key insight is that the measurement of innovation outcomes can be more effective and efficient by linking existing innovation surveys to other available data sources to be used for measuring the business performance. Not bringing the world into innovation surveys but instead bringing innovation surveys to the world opens up a wide array of relevant secondary and tertiary data sources. Nevertheless, the actual availability of the data can be problematic, especially when dealing with commercial data holders. The quality of the data has to be high enough to be suitable for re-use for official statistical purposes, the data has to be accessible, re-use has to be allowed, and it has to be easy to be processed (that is, machine-readable).

The use of tertiary data – 'big data' that is generated as a side effect from business processes – is a promising development and could display some unique features vis-a-vis the use of traditional secondary data (e.g., it allows the [near] real-time measurement of business processes). However, in the absence of common classifications, definitions and formats, big data is difficult to harmonise and therefore at the moment still difficult to translate into existing statistical structures. In contrast, the structure of secondary data is quite similar to primary data. An additional advantage vis-à-vis primary (and tertiary) data is that secondary data (e.g., administrative registers) often cover the whole target population.

As far as the Oslo Manual is concerned, discussing the options for using secondary and tertiary data for measuring the innovation outcome could be beneficial to users. Of course, any practical application of the techniques discussed in this paper will be case (or country) specific and this should prevent from drawing any kind of general recommendation. On the other hand, a reference to existing policy documents that already describe the promises and pitfalls of the use of big data in official statistics could be helpfully included in the Oslo Manual (UNECE, 2013) (Cervera, et al., 2014; Eurostat, 2014) (OECD, 2015).

If any guideline should be given in the Oslo Manual about the measurement of the innovation outcome, it should focus on the (re)use of administrative data. Administrative data could be directly collected from target units (enterprises) or indirectly via public or private data aggregators. The direct (hence primary) data collection of administrative data can be done in the traditional way, by surveys. A promising option to explore is to automate data collection by linking directly to business information systems of individual enterprises. This method is already being used by several tax authorities, and it enables the collection of detailed information (also on the level of individual innovation projects) on business performance without increasing the administrative burden from enterprises.

Where, like in the European Statistical System, an integrated production of official business statistics is consistently implemented at national or international level, the use of primary data on business performance can be recommended in order to derive some innovation outcome indicators. However, where structural business statistics are not produced on a regular basis or cannot easily accessible for the institution managing the innovation survey data, it makes little sense to start collecting outcome data which have already been collected by other aggregators.

In general terms, private institutions running innovation surveys and looking for business performance microdata to be matched in order to produce outcome indicators could find useful to rely on commercially available data. On the other hand, public institutions which own innovation microdata but do not have access to business performance microdata (as well as, NSOs experimenting alternative sourcing) will be quite conscious of the availability issues with the re-use of data from commercial aggregators (including the lack of transparency and the difficulties to harmonise data). In this respect, they could be interested to focus on the re-use of administrative data from public sector organisations.

The use of administrative data from public sector bodies (such as register data) has already a long history in statistics. However, what is new is that the register data is not only used as a basis for population frames but for the actual content of the data. Although the quality aspects of the reuse of administrative data have already been well-described in existing policy documents (Eurostat, 2003) the actual linking with these secondary data sources has not yet been very well covered by the statistical literature (at least, not with specific reference to the application in statistical production).

We therefore suggest to describe in the Oslo Manual strengths and weaknesses in linking to secondary data sources in order to establish the linkage between innovation and outcome.

The integration of other data sources in the statistical infrastructure of a NSO would at least require an information system that could be able to identify and link population units across different internal and external data sets. This does not necessarily require a centralized register but at least a series of compatible registers. In the case of enterprises this requires a unique identification numbering system managed by business registers and used by each of the statistics included in micro-data linking programs. Proper micro integration – that provides far more reliable results from the combination of data sources – is only possible in a fully integrated statistical infrastructure.

The statistical infrastructure that is needed for cross-country linkage of micro-data is already mainly covered in the EU FRIBS program that could be identified as a global best practice. Some parts of the requirements for the infrastrucure are already given in the Sturgeon report (Sturgeon, 2014).

Depending on the type of input data, for the actual linkage two basic options have been described in this paper: record linking and statistical matching.

Statistical matching can be used to combine innovation data (e.g., CIS micro data) with outcome data (e.g, micro data on firm performance) into a merged dataset of synthetic records. This data set could then be used for policy making and research purposes. However one can never be sure to what extent the statistal matching or micro aggregation influences the eventual results of the analyses of the linked datasets. The biggest drawback is the loss of flexibility. The data matching is tailored towards the specific variables that are being combined. This means that the re-use of the generated dataset with generated units for other researches is problematic. A major advantage from statistitical matching vis-à-vis record linking is that the output units do not refer to real-world entities, hence there are much less privacy concerns.

Statistical matching cannot be applied when there is an overlap between the two data sets that are combined. In the case of public registers the coverage of the population is usually nearly complete, Consequently for matching a set of innovation micro data with register data statistical matching cannot be used. Statistical matching is most suitable for incidental research purposes on distinct datasets (that do however have to be harmonized to a certain extent before they can be combined). An integrated statistical infrastructure needs to be built on record linkage.

However a major disadvantage from record linking is that it has inherent privacy and security risks. As it has been discussed above, the two basic measures to protect privacy are to limit access to the data or to conceal the data. Both are consistently used by NSOs when allowing researchers/users to access innovation microdata sets.

In practical terms, the main shortcoming which could be mentioned when proposing the introduction in the Oslo Manual of a specific section on data linking/matching for producing innovation outcome indicators, is the shortage of experience in this area, and related evidence of its feasibility. Moreover, single experiences cannot be diffused in different national and institutional contexts without a long and complex process of adaptation.

Nevertheless, a key point has to be introduced into the Oslo Manual revision debate: that of the distinction between the measurement of short-term and long-term outcomes of the innovation activity. So far, it has been tacitly agreed that (official) statistical production should focus only on effects, i.e. short-term outcomes, leaving any estimation of the longer-term outcomes of business innovation to the analyses of researchers and academicians. Currently, the increasing availability of statistical and digital infrastructures by the NSO (and, more in general by the institutions running innovation surveys) could allow for experimenting the production of innovation outcome indicators – as a result of a systematic matching of innovation and business performance databases – as a key role of the official statistical institutions.

5 References

Al, P., & Thijssen, J. (2003). Bespiegelingen over het waarom, de mogelijkhedenen beperkingen van micro-integratie in de sociale statistieken. In J. Nobel, M. Algera, M. Biemans, & P. v. Laan, Gedacht en gemeten (pp. 112-122). Voorburg/Heerlen: Statistics Netherlands.

Augustyniak, L. (2015, July). Big Data. Zettabyes and yottabytes of information. Commission en Direct, p. 42.

Bakker, B. (2011). Micro integration. The Hague: CBS Statistics Netherlands.

Boer, Y. v., Arendsen, R., & Pieterson, W. (2016). In search of information: Investigating source and channel choices inbusiness-to-government service interactions. Government Information Quarterly, 33(1), 40-52.

Brennenraedts, R., & te Velde, R. (2012). Internet as data source. Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering. Brussels: European Commission, Communications Networks, Content & Technology Directorate-General.

Buelens, B., Boonstra, H., Brakel, J. v., & Daas, P. (2012). Shifting paradigms in official statistics. From design-based to model-based to algorithmic inference. The Hague/Heerlen: CBS Statistics Netherlands.

Buterin, V. (2014, 2 9). A Next-Generation Smart Contracts and Decentralized Application Platform. Opgehaald van Github: https://github.com/ethereum/wiki/wiki/White-Paper

Cervera, J. L., Votta, P., Fazio, D., Scannapieco, M., Brennenraedts, R., & van der Vorst, T. (2014). Big data in official statistics. Technical workshop report. Rome: Eurostat/ESS.

Defays, D. (1997). Protecting micro data by micro-aggregation: the experience in Eurostat. Qüestiió, 21(1), 221-231.

Dias, C. (2015). Statistical Matching and Data Linking. presentation at Eurostat, 29-30 November. Luxembourg.

Donnellan, B. M., & Lucas, R. E. (2013). Secondary Data Analysis. In T. D. Little, The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis (pp. 665-677). Oxford: Oxford University Press.

Eurostat. (2003). Quality assessment of administrative data for statisticical purposes. Luxembourg: Eurostat.

Eurostat. (2011). European statistics code of practice. For the national and community statistical authorities (28 September 2011 ed.). Luxembourg: European Commission.

Eurostat. (2012). The Community Innovation Survey 2012. The harmonised survey questionnaire. Luxembourg: Eurostat.

Eurostat. (2013). Statistical matching: a model based approach for data integration. Luxembourg: Eurostat.

Eurostat. (2014). The ESS Vision 2020. Opgehaald van http://ec.europa.eu/eurostat/documents/10186/756730/ESS-Vision-2020.pdf/

Eurostat. (2015a). Framework Regulation Integrating Business Statistics (FRIBS) Roadmap - Update. Luxembourg: Eurostat.

Eurostat. (2015b). Summary report on the open public consultations on FRIBS. Luxembourg: Eurostat.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64, 1183-1210.

Gault, F. (2014). Where are innovation indicators, and their applications, going? Maastricht: UNU-MERIT.

Hagsten, E., Polder, M., Bartelsman, E., & Kotnik, P. (2013). The multifaceted nature of ICT. Final report of the ESSnet on Linking of Microdata to Analyse ICT Impact. Stockholm: Statistics Sweden.

Kloek, W., & Vâju, S. (2013). The use of administrative data in integrated statistics. Brussels: NTTS2013.

Lamarche, P., & Pérez-Duarte, S. (2015). Microaggregation for the masses: non-confidential enterprise-level data for analytical and research purposes. New Techniques and Technologies for Statistics. Brussels: Eurostat.

Laux, R., & Radermacher, W. (2009). Building Confidence in the Use of Administrative Data for Statistical Purposes. Durban: International Statistical Institute.

Montoye, Y.-A. d., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobilityUnique in the Crowd: The privacy bounds of human mobility. (3). doi:10.1038/srep01376

OECD. (2005). Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data (3rd ed.). Paris: OECD.

OECD. (2015a). Data-driven innovation. Big data for growth and well-being. Paris: OECD.

OECD. (2015b). Scoping the Third Revision of the Oslo Manual. Paris: OECD.

OECD. (2016). Rethinking Tax Services: The Changing Role of Tax Service Providers in SME Tax Compliance. Paris: OECD. doi:10.1787/9789264256200-en

Provost, F., & Kohavi, R. (1998). On Applied Research in Machine Learning. Machine Learning, 30(2), 127-132.

Rammer, C., Schubert, T., Hünermund, P., Köhler, M., Iferd, Y., & Peters, B. (2016). Dokumentation zur Innovationserhebung 2015. Mannheim and Karlsruhe: Zentrum für Europäische Wirtschaftsforschung (ZEW).

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Stat Sci., 25(1), 1-21.

Sturgeon, T. J. (2014). Global Value Chains and Economic Globalization. Towards a New Measurement Framework. Boston, MA: Massachusetts Institute of Technology.

te Velde, R. A. (2007). Quality of Public Sector Information. ePSIplus Thematic Meeting (Pricing 1). Helsinki: ePSIplus.

Torra, V., & Navarro-Arribas, G. (2015). Data Privacy: A Survey of Results. In G. Navarro-Arribus, & V. Torra (Red.), Advanced Research in Data Privacy (pp. 27-37). Cham: Springer International Publishing.

UNECE. (2013). What does "big data" mean for official statistics? Geneva: UNECE.

Wallgren, A., & Wallgren, B. (2011). To understand the Possibilities of Administrative Data you must change your Statistical Paradigm! Section on Survey Research Methods (pp. 357-365). Miami: Joint Statistical Meetings.

Wilhelmsen, L. (2014). Assessing a combined survey strategy and the impact of response rate on the measurement of innovation activity in Norway. The Statistics Newletter, 60(January), 3-5.

Wilhelmsen, L. (2015). A Schumpeterian Problem -- Making strict Rules for Measuring Fuzzy Concepts. Oslo: Statistics Norway.