4. Barriers to data access and matching in Europe: concluding remarks

This Blueprint so far has investigated the extent to which a wide range of competitiveness indicators, especially those that are built from micro-data and that we have defined as bottom-up indicators, can be computed for EU countries and what data is actually accessible for researchers. In chapter 2, we highlighted issues at the level of individual countries, while in chapter 3, we focused on the challenges of using micro-data to construct indicators of competitiveness across countries. In this chapter, we pick up on the main conclusions emerging from chapters 2 and 3 (in sections 4.1 and 4.2, respectively). Building on these considerations, in the next chapter we offer some policy recommendations.

4.1 Issues regarding the availability of data at country level

The availability of an indicator of competitiveness depends on different factors. In the MAPCOMPETE data mapping exercise (see chapter 2), we distinguish between factors that determine the computability of an indicator and factors that influence accessibility. By computability we mean the quality of data and the length of time coverage. Computability of an indicator relies mainly on data existence and the possibility to merge data from different sources, if necessary. The accessibility of data depends on the rules of access and their clarity. As part of the MAPCOMPETE data mapping exercise, statistical institutes of EU member states were approached to collect information on micro-data availability. Project participants surveyed several bottom-up competitiveness indicators – firms' productivity, dynamics, international activities, R&D activities and some other features – with respect to computability and accessibility.

4.1.1 Availability of data for statistical/research purposes

MAPCOMPETE participants surveyed several bottom-up competitiveness indicators, which are based on basic information about enterprises, such as year of establishment, number of employees or financial statement and balance sheet items. Although such information is usually collected by national authorities for administrative purposes, our findings on the availability of this data present a mixed picture.

We find that those indicators that require the use of basic balance sheet data (eg labour productivity, TFP) – along with trade indicators – are the most computable among the bottom-up indicators we surveyed, but there are country-specific problems. Also, bottom-up indicators on firm dynamics, which are based on data about company entries and exits, are poorly computable for several member states. In some cases the information needed is available, but only for a subset of enterprises or for a limited time period.

Much of this heterogeneity can be explained by the fact that countries report various databases as the best possible source of information on firm dynamics, balance sheet and financial statement items. There are NSIs that report survey data as the best possible source of information, while others indicate that administrative databases are available for statistical use.

Our findings are consistent with the findings of a recent ESSnet project. The ESSnet Admin Data project[75] examined the use of administrative and accounts data for producing national statistics. The project outcomes show that both legislation and existing practices regarding the use of administrative data differ in different EU member states[76]. They highlight the possibility to improve the quality of business statistics and to reduce the administrative burden on enterprises by finding common ways for using administrative data. It is also stated that relevant administrative data is available to a greater extent than is actually used. In some countries, administrative data is only used as a sampling framework, or for imputation and validation, while NSIs compute national statistics using survey data.

In most member states, national legislation supports the use of administrative data

75. http://essnet.admindata.eu/.
76. Costanzo, L. (2013) Report to Eurostat on the "Overview of Existing Practices", Admin Data, Work Package 1. http://essnet.admindata.eu/Document/GetFile?objectId=5995.

for statistical purposes – under different confidentiality restrictions – and provides special rights for the NSIs to access these sources. However, the ESSnet Admin Data project identified several factors that hamper the effective use of administrative sources. First, legislation that requires the use of administrative data whenever possible is rare (exceptions are Finland and the Netherlands). As a consequence, NSIs are not motivated to make investments in order to fully exploit administrative data. They use such data, but only if it can be used with minor adjustments as part of existing practices.

Second, most countries lack a coherent and comprehensive framework for collecting, storing and providing access to collected data. Different production units of NSIs perform admin-data related tasks separately, thus the use of administrative data is based on *ad-hoc* agreements with limited scope between the NSIs' production units and the data holders. There are, however, positive examples: Portugal replaced all surveys of Structural Business Statistics with one new data-collection system for administrative and statistical use, while Bulgaria introduced a single entry point for reporting fiscal and statistical information.

Third, cooperation between admin-data holders and NSIs is weak or difficult in several countries, partly because of the lack of legislation establishing the corresponding duties of data holders. In most countries, NSIs have no impact on the design of administrative data collection and authorities do not have to consult NSIs when introducing changes to data collection practices.

These aspects have been addressed in a amendment to Regulation (EC) No 223/2009 – being finalised at the time of writing[77] – which aims at establishing a legal framework for more extensive use of administrative data sources for the production of European statistics without increasing the burden on respondents, NSIs and other national authorities. NSIs should be involved, to the extent necessary, in decisions about the design, development and discontinuation of administrative records that could be used in the production of statistical data. NSIs should also coordinate relevant standardisation activities and receive metadata on administrative data extracted for statistical purposes. Free and timely access to administrative records should be granted to NSIs, other national authorities and Eurostat, but only within their own respective public administrative system and to the extent necessary for the development, production and dissemination of European statistics.

77. See footnote 57.

4.1.2 Legal and administrative constraints of access to micro-level data

The MAPCOMPETE data mapping exercise revealed substantial differences between EU member states in terms of the accessibility of micro-level information needed to compute the surveyed competitiveness indicators. We observe that there are countries for which many bottom-up indicators have a relatively high level of computability, meaning that the required information exists in some meaningful format at the local statistical authorities, but micro-data access is not allowed for outside users.

Legal barriers related to confidentiality

While the rules of micro-data access are not clearly specified in several countries, it is clear that confidentiality restrictions substantially differ in different member states. The common feature of national laws is that they oblige institutions collecting personal or firm-level data to guarantee the anonymity of respondents. However, various definitions of confidential data and different approaches to data protection are present. Research entities have the option to access personal data in the majority of countries, but there are significant differences in national confidentiality restrictions regarding the transmission of data from the collecting institution to other entities[78]. Some member states do not allow the transmission of certain confidential data, or the implementation is problematic.

Importantly, regulations concerning Eurostat itself also differ in different member states: Eurostat can't access confidential data from some countries.

The new EU statistical law[79] emphasises the importance of the availability of confidential data within the ESS network. It states that the transmission of confidential data between ESS partners may take place *"provided that this transmission is necessary for the efficient development, production, and dissemination of European Statistics or for increasing the quality of European statistics"*. The access to confidential data for scientific purposes also requires the approval of the national authorities which provide the data. However, our experience suggests that despite the legislative underpinning, there are several factors that hinder the research use of micro-data, and the exact methods, rules and conditions of access are still to be developed in many member states.

78. Ichim D., Franconi L. Strategies to achieve SDC harmonisation at European level: multiple countries, multiple files, multiple surveys, http://neon.vb.cbs.nl/casc/..%5Ccasc%5CESSnet%5Ccomparable%20dissemination%20v-1.pdf
79. Eurostat, Legal Framework for European Statistics - The Statistical Law, 2010 Edition, http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-31-09-254/EN/KS-31-09-254-EN.PDF

The mapping of micro-level information also highlights the fact that different types of data are treated differently. In some EU member states, different regulations apply to different databases. Databases with the full population compiled by National Statistics Institute of Italy are not accessible to researchers, who can only access descriptive statistics upon request, but micro-data stemming from surveys is available. In the Czech Republic, business register data can be accessed relatively easily, while for other types of data, such as custom data and FATS data, conditions are more stringent. Malta allows access to firm-level information for research purposes, except for data on foreign ownership and capital. In Latvia, data is available upon request, except for data on trade by destination and product, which is confidential.

Our results show that in general there are stricter regulations on registry-type data and on databases that have full coverage over the observed population. Survey type data, especially data from harmonised surveys like CIS, is usually easier to access. Our findings on individual-level trade data are mixed, since these databases include information both from administrative sources (ExtraStat) and from a harmonised survey (IntraStat).

A distinction in confidentiality restrictions is particularly important when we consider the potential use of bottom-up indicators that are based on information obtained from different sources in different countries. For instance, firm entry and exit information and balance sheet data are obtained from administrative sources in some countries, while others conduct surveys to collect the information. Consequently, the computability and accessibility of bottom-up indicators based on these data is likely to differ in different countries and a harmonised approach to confidentiality protection is hard to achieve.

It is worth mentioning that Eurostat provides access for scientific purposes to certain European survey data[80] including the Labour Force Survey and the Community Innovation Survey. Recognised research entities conditional on the approval of their research proposal might access micro-data anonymised by Eurostat on electronic devices or non-anonymised data in Eurostat's 'safe centre'. Currently, Eurostat negotiates on the possible dissemination of the micro-data on a case-by-case basis and proposes a unique anonymisation methodology to all member states. Member states might refuse Eurostat's proposal if it conflicts with national legislation, and thus micro-data will not be available for all member states[81].

80. Comission Regulation 831/2002 specifies the surveys and the rules of access.

81. Ichim D., Franconi L. Strategies to achieve SDC harmonisation at European level: multiple countries, multiple files, multiple surveys, http://neon.vb.cbs.nl/casc/..%5Ccasc%5CESSnet%5Ccomparable%20dissemination%20v-1.pdf

Practical (technical) constraints on accessibility

We observe that in addition to national legislation, the internal regulations of data-collecting institutions and practical constraints also affect the accessibility of micro-data. In Romania, practical barriers hinder the accessibility of the databases compiled by the NSO: a safe environment for data security is at the time of writing not yet in place. Part of the variation in these matters can be explained by the fact that the increased demand for micro-data is a relatively new phenomenon. The resources available to NSIs for disclosure control, and their prior experience in the field, might influence the speed and direction of adaption. The development of new statistical disclosure methods needed to provide access to micro-data might be hindered by organisational, methodological and software problems.

Our results show that currently, at the national level, the most commonly used method to provide access to micro-data is the release of scientific use files. In case of research use files, statistical disclosure methods and restrictions on access and use – eg license or access agreements – are applied simultaneously[82]. Our data mapping exercise shows that several NSIs provide access to micro-data in data laboratories. Data laboratories allow researchers to use more identifiable data under strict conditions. In most cases, users are legally obliged to keep the data confidential, and are subject to close supervision and output checking. Since setting up a data laboratory takes time and resources, there are countries where this form of micro-data access is not yet available. Remote execution is also possible in a few member states. Note that the cost of operating a data laboratory or remote access services significantly increases with the number of users, mostly because output checking is completely manual in almost all of the member states. Consequently, even in the countries where the NSI already provides access to micro-data, revision of data protection practices will be inevitable in the near future.

4.1.3 Non-legal barriers

Issues with metadata

Having basic information about datasets in advance is a very important factor that might affect the success of a research project. Researchers need to have detailed

82. Eurostat, Handbook on statistical disclosure control (January 2010) http://unstats.un.org/unsd/EconStatKB/Attachment474.aspx

information on the available datasets including the identity of the owner of the data, the exact content, the quality of data and the rules of access. These pieces of information are necessary to decide whether the dataset is suitable to their needs and whether they apply for access.

International standards already exist for the international exchange of metadata. Statistical Data and Metadata Exchange (SDMX), an initiative sponsored by the Bank for International Settlements, ECB, Eurostat, International Monetary Fund, OECD, United Nations and the World Bank, aims to provide standards for the exchange of statistical information (eg formats for data and metadata, content guidelines, IT standards)[83]. Particularly for Europe, the European Commission has set up a recommendation 'on reference metadata for the European Statistical System'[84], which refers to the European Statistics Code of Practice[85] and is based on the SDMX framework.

While ESMS Metadata files for all of the statistics published by Eurostat are provided – and other international organisations also provide structured metadata on their statistics – our experience shows that there is still a big hole in the information on data. ESMS metadata files present useful information on methodologies, quality and the statistical production processes in general, but usually provide very little information on the link between the aggregate indicator and micro-data used to compute the given indicator. Also, country-specific information on survey and sampling design is often sketchy. We made use of the information provided in ESMS Metadata files when mapping the readily-available aggregate indicators, but we found that in order to be able to assess the strengths and weaknesses of these indicators to improve their quality or to propose new ones, much more information on the available national micro-data would be needed.

Gathering comprehensive information on micro-data available in EU member states proved to be a challenging and time-consuming task. The amount and structure of information available on the websites of NSIs and other national data providers is very different in different countries. It is usually insufficient to fill the MAPCOMPETE MetaDatabase and it is definitely insufficient to plan a research project. In many cases, researchers obtain information on given datasets from scientific publications or

83. SDMX (2009), Content-Oriented Guidelines, Statistical Data and Metadata eXchange. Vale, S. (2009), Generic Statistical Business Process Model, Version 4.0 – April 2009, UNECE Secretariat.
84. See European Commission (2009), Commission recommendation of 23 June 2009 on reference metadata for the European Statistical System, Official Journal of the European Union L 168/50, 50-55.
85. Eurostat (2011), European Statistics Code of Practice for the National and Community Statistical Authorities, Eurostat, European Statistical System, Luxembourg.

through informal channels, which are burdensome and usually result in incomplete information. Also, when conducting cross-country comparative research or research that requires the use of information from more than one source, researchers have to search through several websites and publications, each with different metadata structure and information content.

Since in MAPCOMPETE we collected a huge amount of information in a systematic manner, we tried to directly contact staff within the NSIs in all the EU28 countries to gather the relevant information. After a few months of the project, it became apparent that this was highly complicated, so we decided to gather information by exploiting existing contacts built up in another international project (CompNet) and from other personal contacts. In some cases, these contact persons were able to help us fill in the MAPCOMPETE MetaDatabase and in other cases they referred us to people within the NSI. The fact that in most countries economic databases are collected and handled by more than one institution – the NSI and the national central bank (and sometimes other institutions) both collect data in most cases – made it even harder to obtain the required information. Also, smaller countries and newer EU members tend to have less experience in handling requests for micro-data access, and consequently are usually less prepared to provide systematic information on existing data.

The experience we gained during the data-gathering process shows that the availability of information on the data is at least as important as the availability of data itself. Performing EU-wide research projects on competitiveness or designing new indicators is not feasible without easily available, comprehensive information on national micro-data. This is why the MAPCOMPETE MetaDatabase is especially useful for future research on measures of competitiveness. Furthermore, it serves as a basis for suggestions for possible improvements to data sources, treatment of data, conditions of access etc. It might promote quality research by providing detailed information on the accessibility and availability of data related to the measurement of competitiveness. However, the MAPCOMPETE MetaDatabase is only a snapshot of competitiveness-related data. A regularly updated, structured, easily available and comprehensive meta database on national micro-data – that might include the experience of other researchers working with the data – might substantially increase the efficiency of international research projects.

Issues related to the nationality of the data user

As part of establishing the European research space, conducting research and analysis on the basis of foreign data becomes important. Several specific problems arise in

terms of foreign access to datasets located in countries other than the nationality of the researcher. First, in some countries, such as Belgium, Denmark, Hungary and the United Kingdom, access to micro-data is allowed only to researchers who are citizens of the country of the data provider or affiliated with a national institution. Second, language barriers are obviously a serious burden, since in many countries information is provided only in the national language, but one that can be solved by simply offering data description and variables in English. Several NSIs have made a great deal of progress in this respect, including metadata provision in English. Third, the provision of data on site might not be a burden for locals, but can be very costly for foreign researchers. Hence, setting up secure remote access – such as is available in Finland, France, Germany and Sweden – would be an important step. Finally, making access by foreigners easier by appointing an English-speaking specialist could indeed facilitate European research integration.

Unclear rules of access

When mapping the accessibility of data, we faced the obstacle that it is often challenging to obtain precise information on the conditions of access to confidential data. Information on the accreditation process, statistical disclosure control methods applied and the practical details of access is usually not clearly specified on the website of the data provider or at any other publicly-available source. We found that one had to contact the data provider directly in order to clear up the details and to find out if access to the data is possible and under what conditions.

Our results show that there are substantial differences between countries in terms of the clarity of rules of access. In many countries there is some settled, formal procedure of applying for access (eg Denmark, Finland, France, Netherlands, Slovenia and Sweden) while other countries are less advanced in this respect and handle requests on a case-by-case basis. However, regardless of the sophistication of the application procedure, in most cases, it is required to present a research project which needs to be approved. This approval creates room for discretionary decision-making and informality which might differ from country to country, but is really difficult to assess.

The approval procedure might be more problematic when the data provider does not perform output checking itself, but it is the researcher's responsibility to protect the confidentiality of data. If data protection is delegated to the researchers then the cooperation strongly relies on trust between the data provider and the researcher, and it might be hard to define exact criteria.

Truncated data

In many cases, micro-data is provided in truncated form; that is it is made available with less information than the original source, in order to prevent the risk of disclosure (sensitivity) and for cost reasons. For the purposes of our discussion, this aspect is related to accessibility, but it can affect computability when it prevents the merging of different datasets.

Sensitivity truncation

Several statistical disclosure methods used to protect the confidentiality of data lead to a loss of information and might affect the quality of analysis carried out on the data. Let us first present key obstacles and make suggestions for their treatment (for details and a broad discussion, see Hundepool *et al*, 2010). According to statistical best practice, this implies *"first a definition of possible situations at risk (disclosure scenarios) and second, a proper definition of the 'risk' in order to quantify the phenomenon (risk assessment)"* (Hundepool *et al*, 2010, p. 30).

In this chapter, we identify four issues that matter for practitioners:

1. Sensitivity of information on selected firms;
2. Recoding data into broader categories;
3. Removing or modifying variables;
4. Other disclosure measures.

The first issue is related to the sensitivity issues of aggregated data. In some sectors, size categories or regions, there are only very few firms. Aggregating data on them would imply that in some categories only one or very few firms would feature and hence, their individual data would not be protected. To avoid this scenario, most statistics institutions and central banks or research outlets protect confidentiality by setting up compulsory aggregation rules. Typical rules include a minimum number of firms per aggregated band (this ranges between 4 and 9, in our experience) and maybe other controls such as market share of the top 5 firms in the aggregate.

The second topic is a more general solution to keep identification impossible. This entails aggregating some existing firm categories such as industry or location address to protect the identity of firms. This process is especially useful in smaller countries where some regions or industries might include only a few firms, even if they are not large. Examples include merging four-digit industry codes into two-digit codes, merging

municipalities or NUTS3 regions into NUTS2 regions, or replacing employment data with firm size brackets.

Third, authorities might remove or replace variables. This might include the deletion of variables that would allow identification – this happens when some activity occurs rarely or is carried out by only a few firms. This might include balance-sheet items, such as subsidies, or some research activities in an innovation survey.

Another option to prevent identification in general, and merging of datasets, in particular, is masking. This approach is divided into two categories depending on their effect on the original data: perturbative and non-perturbative masking methods. Perturbation implies the multiplication of all values by a random variable of unit expected value and a small but significant variance. This implies that say, sales values would be altered by a few percent without affecting any statistical relationship (given the unit expected value). Other options include rounding or truncation. In these cases identification or linking of the data to other data sources would be impossible or difficult because of the lack of exact matching (for more details, see Willenborg and de Waal, 2001).

Importantly, researchers can often access sensitive information in, for example, the research lab, but there are strict rules for the information available outside the safe environment. Apart from these more common issues, authorities might apply individual controls or ask for a list of descriptive statistics to control the process. Statistics offices will often ask researchers to submit all relevant documentation – including programme code files, and descriptive tables for output checking before releasing results.

Finally, note that in some cases an extreme application of this sensitivity approach is applied: individual data is aggregated right after data collection. In this scenario, firms are clustered by industry, location, size and only aggregate information is released. While this may indeed provide security, it washes out important features of observations that may be important for research.

Dataset reduction for cost saving

Another factor that might reduce the scope of available datasets is cost saving. Every aspect of a dataset – number of variables, dimensionality and frequency of observations – will generate additional costs, mainly in terms of attention. Supervisors need to spend time on organisation of dataset management, cleaning and provision,

and the costs of these will depend on the size and complexity of the data at hand. Saving resources and reducing administrative burdens are important in an era when NSI budgets are often being cut. As a result, aggregation and truncation of raw data are often carried out not for sensitivity but for cost purposes.

One such practice is aggregation of some part of the dataset. Transaction-level data might be aggregated into annual aggregates. For instance, foreign trade is often registered at a very fine transaction level, but available data is mostly at annual aggregate level. Several variables might be deleted in order to avoid spending the time that would be required for consideration of sensitivity issues.

Finally, another approach is exclusion of small firms. Dropping firms with fewer than five employees could reduce the size of a dataset by 80-90 percent, while retaining 95 percent of value added. However, such an exercise will limit analysis and understanding of important issues, such as entrepreneurship and firm dynamics.

An important aspect of dataset reduction for cost saving reasons is European/international harmonisation. Comparing statistics computed on the whole dataset or on firms with more than 10 employees might yield rather different results (for an application for exporters, see Békés *et al*, 2011).

4.2 Accessibility and matching of data from different countries

As we argued in chapter 3, data matching opens up rich and novel research opportunities, especially when micro-level datasets are concerned. Existing micro-level data in European countries has significant potential in terms of record linkage and matching, including also commercial data and Big Data. Data matching and issues of matchability have considerably gained in importance in recent years. One reason for this lies in the increased accessibility of micro-level datasets and in the desire of researchers to merge these datasets within and between countries in order to increase the research potential of the data. There has also been significant progress on technical issues, not least driven by the rapid development of computer technology and data storage.

The issue of data matching and matchability is of course not confined to the social sciences, but the recent economic crisis has made clear that economists require high-quality data, especially at the micro level, that is comparable across countries, in order to examine cross-country differences in competitiveness. However, comparable micro-data at the firm level in different EU countries is so far only available for some topics,

most of which are not directly relevant for competitiveness (notable exceptions are the Community Innovation Survey, the International Sourcing Survey or the EFIGE survey). These comparable micro-level datasets are, however, all based on sample surveys.

The huge potential of administrative data, which is already leveraged in many countries, is still waiting to be fully realised (see Agafitei and Vaju, 2013, for instance). There are, however, some serious endeavours in this direction, mainly based on the ESSnet projects and on the Framework Regulation for Integrating Business Statistics (FRIBS, see section 3.2). These projects are of special importance because they are concerned with administrative data within the EU, which is of high quality. Any step towards making these data more comparable and accessible is more than welcome by researchers and policymakers. Therefore, ensuring the availability of such data should be a priority for the European Commission because this would ensure vastly improved analysis of cross-country differences in competitiveness, and of labour market issues and related fields.

The most serious obstacles to matching micro-level data from different countries are still legal restrictions preventing data from being matched, because privacy and confidentiality are at stake. However, there is some activity in this area, namely within projects to evaluate the potential of analysing micro-level data without directly accessing the data.

There are also obstacles to data matching within countries (see the KombiFiD example from Germany). This holds especially true if the datasets to be matched are held by different data providers, eg statistical offices, central banks, employment agencies or private data providers. However, progress has been made in this regard in recent years.

Important steps to overcome the problem of data comparability between countries, particularly with regard to cross-country analyses of competitiveness, have been taken, for instance by the EFIGE project providing comparable firm-level data for 15,000 firms from seven EU countries. The ECB's CompNet project is following suit. However, these two projects can only be regarded as first tentative steps towards data that can be used for cross-country analyses in the field of competitiveness, and that is highly useful for policymakers.

Overall, much has been achieved in the field of data matching within Europe in recent years, but the universe of cross-country and matched datasets is still sparsely populated and quite heterogeneous, with potential for improvement. Because of the

ever-increasing need for high-quality datasets that can be used to inform policymakers, much more needs to be done. Cooperation between data providers within and in different countries is key, as is the reduction of red tape. Comparative analysis of competitiveness in different countries is ultimately only possible if comparable (micro) data exists in different countries or if data can be harmonised and made accessible to researchers. Ensuring the availability of such data should be a priority for the European Commission, because it would enable vastly improved analysis of policy-relevant issues.