www.makswell.eu

Horizon 2020 - Research and Innovation Framework Programme
Call: H2020-SC6-CO-CREATION-2017
Coordination and support actions (Coordinating actions)

Grant Agreement Number 770643

Work Package 5

Pilot study for integrated frameworks at different territorial levels and measurements for policy making

Deliverable 5.1

Reflection Paper

"Future research needs in terms of statistical methodologies and new data"
May 2018
Istat

Authors:
This work has been coordinated by Tommaso Rondinella at Istat and results from the active contribution of all partners who revised an initially proposed text during two rounds of amendments

This project has received funding from the European Union's Horizon 2020 research and innovation programme

Deliverable 5.1

Reflection Paper

“Future research needs in terms of statistical methodologies and new data”
Early reflection paper to define future pathways with respect of new EU Framework Programme (FP9)

Summary

On the basis of the analysis of the major transformations affecting official statistics and the challenges that National statistical offices and data producers are called to face, the paper presents a set of recommendations to the European commission for shaping the forthcoming 9th Framework Programme which mainly refer to activities dedicated to:

- quality and timely data for the full implementation of the 2030 agenda for sustainable development:
- developing methodologies for big data treatment and production of new indicators;
- enhancing integration between surveys, administrative data and new sources;
- a shared big data quality framework;
- availability of evidence-based policy tools at different territorial level;
- extension of open data platforms to the whole public administration;
- extending statistical literacy through formal and informal education.

Index

1. Rationale and background

The evolution of the global landscape is the product of social and economic transformations driven by digitalization, globalization, artificial intelligence, and a higher degree of interconnections among firms, households and territories. Their interpretation and governance represent a big challenge that requires new metrics: indicators able to define new conceptual frameworks for an enhanced evidence-based policy making. “Indicators can be considered as the road signs of policy making” (Eurostat, 2017a).

At the same time, new technologies and digitalization processes have dramatically increased the potential for data production, collection and analysis, and it is now common to talk about the “**data deluge**” to refer to the huge amount of information we face on a daily basis.

It is a transformation happening only partly within the perimeter of official statistics. Most of it derives from the explosion of new sources made available through the use of ICT and the internet: sensors of any kind, digitalization of business activities, internet and social media are all producing an incredible wealth of data (so-called **big data**). Technological progress is leading to an increase in the number of data producers and the range of measured phenomena, and, in parallel, an increasing demand for data from all parts of society (Data Revolution Group, 2014).

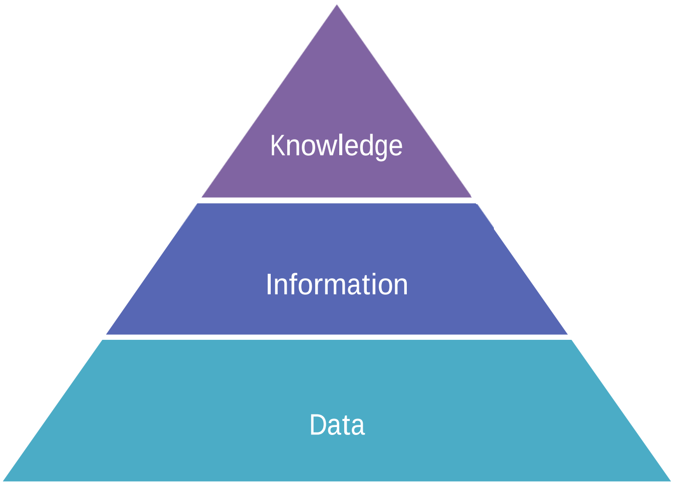
New technologies allow for the use of a number of new sources, whether big data or administrative sources, and their integration with the usual surveys. The so-called ***modernization*** of statistical production envisages an overhaul of the traditional production model, based mainly on surveys for the direct acquisition of data from citizens and businesses, towards a model based on the integration of individual data from a plurality of sources (surveys, administrative archives and new sources) and the creation of statistical registers of individuals and households, economic units or geographical units (ECOSOC, 2016). The potential of new information is huge, so that new demands for more cross-cutting, timely, disaggregated, geo-referenced data are continuously emerging.

New sources are expected to grow further including also the so-called **smart statistics** (Eurostat, 2016), where data capturing, processing and analysis will be controlled through artificial intelligence processes. The artificial control of the entire ecosystem will lead to the production of “relevant statistics, largely instantly and in an automated way”. Artificial intelligence has become an area of strategic importance and a key driver of economic development within the European Digital Single Market strategy and data production is bound to be one of the sectors of application.

The broad use of new sources raises the issue of quality control. Official statistics is characterized by high quality standards, including definitions, classifications, production methodologies, punctuality, etc., as indicated by the European Code of Practice (ESSC, 2017). A **quality framework** for traditional sources is nowadays well-established. Yet, it has been built having in mind the traditional production process of well-structured data, either sample surveys or administrative data. But reflections on the quality level are now on the agenda of the statistical community when addressing integrated registers or new sources as they are easily affected by selection and measurement biases. In these cases the quality framework may differ from the one actually used by National Statistical Institutes (NSIs) due to the different nature of the sources, or to often innovative and experimental data treatments. A preliminary framework was developed at UNECE (2014) building on dimensions and concepts from existing statistical data quality frameworks, and proposing a number of needed extensions. Yet, the mainstreaming of big data quality frameworks is still part of the future agenda of official (and non-official) statistics.

The knowledge **pyramid** and the explicit consideration of the users' point of view are now part of the background for improving statistical quality. The knowledge level, on top of the pyramid, refers to 'framed experience, values, contextual information, expert insight, and grounded intuition that provides an environment and framework for evaluating and incorporating new experiences and information' (Eurostat, 2017b).

Figure 1. The knowledge pyramid.



Source: Eurostat, 2017b

Today's and future **challenges** for official statistics producers are many and substantial.

First of all, NSIs face a data revolution that needs to be managed: legal frameworks, IT tools, methodologies and skills are all aspects to be addressed by both the public and the private sector, and for which research and innovation are needed. Instead of focusing (only) on a "big data or data revolution," it is time to focus on an "**all data evolution**", using data from all traditional and new sources, and providing a deeper and clearer understanding of the problem at hand.

The development and diffusion of new digital technologies have knocked down many obstacles, first and foremost, to the production, storage and analysis of information; other actors, public and private, are now able to collect, process and communicate statistical data as never before. The statistical institutes therefore find themselves competing with **other producers**, who often supply more timely data but respect less stringent quality constraints. To foster trust, it is important to disseminate the message that quality intrinsically characterizes official statistics.

Thirdly, the increasing complexity of modern societies and the multidimensional nature of the phenomena under study (e.g. sustainability, globalization, well-being, social exclusion, the environment, and competitiveness) require a continuous expansion of statistical information to satisfy **new and more specific knowledge needs**, either of a thematic nature (economic, social, environmental, etc.), or of territorial detail (from global phenomena to micro-territorial tendencies), or of type of information produced (aggregated data, microdata, microeconomic studies, composite indicators, visualisations, etc.). One key area is represented by the Sustainable Development Goals (SDGs) indicators. In the document "Transforming our world: the 2030 Agenda for Sustainable Development" (UN, 2015), the United Nations highlighted the need for "quality, accessible, timely and reliable disaggregated data [...] to help with the measurement of progress and to ensure no one is left behind" (paragraph 48).

The urgent need for new data and the availability of innovative, yet not standardized methodologies is leading to the production of experimental statistics. **Experimental** statistics[1] are compiled from new data sources and methods. Examples are the now-casting estimates ("flash estimates") on SILC data providing timelier social statistics on income poverty and inequality, or the use of Wikipedia as a new source to produce statistics on the online visits to UNESCO World Heritage Sites as a measure of their popularity. Other examples are the integrated information on consumption, income and wealth, experimentally merged even if proceeding from different surveys. These new frameworks try to fill the gaps coming from new issues that have to be tackled by public policies. In other words, Eurostat, as other NSIs, is extending its boundaries, upgrading along the pyramid from the information level to the knowledge level.

Data collected, produced and disseminated by official statistics institutions provide a solid and irreplaceable foundation for political decisions (not only in the sphere of sustainability), linking them to the reality of the country. In fact, in recent years, **evidence-based policies** have acquired great importance. Very relevant examples are the sets of indicators adopted by the European Union to support, among others, the Macroeconomic Imbalances Procedure, the Europe 2020 initiative, the Cohesion policies or the Common Agricultural Policy. In these fields, the principle of relevance, i.e. the ability to produce statistics capable of responding to the knowledge needs of institutions, public administrations, the research world, and civil society, becomes particularly important.

This role of the NSIs puts more pressure on the extension of **models** able to produce knowledge. Extended macroeconomic and microeconomic models are needed to gauge the possible impact of policy measures on non-economic phenomena. Improving the ability of the statistical institutes to provide a clear picture showing the relationships between policies and their effects will be an important challenge, especially when policies aim to extend their goals to well-being and sustainability.

Therefore, statistical production must progressively move towards a paradigm of "**statistical service**" (Eurostat, 2012, p 54). NSIs are called to support citizens and policy makers in data use and to tailor information to users' needs, but to still, of course, maintain that independence which is fundamental for guaranteeing public legitimacy and the trust of all social actors.

The added value of produced information requires, on the other hand, that users and citizens have the adequate tools to correctly interpret the information. This is more and more urgent in the "era of fake news", when all citizens should be able to recognise reliable sources and find useful data for interpreting social, economic and environmental phenomena. The activities of **training and promotion of a statistical culture** as a whole are an important opportunity to convey and strengthen the role of official statistics, as they allow us to get in touch with large segments of the population, even those less familiar with statistics.

[1] http://ec.europa.eu/eurostat/web/experimental-statistics

The formal and informal training and education processes must find completion and support in a stable, organic and systematic relationship between the official statistics subjects and the world of school and university, with shared training objectives. The European Master in Official Statistics (EMOS) initiative is a fruitful effort in this direction that is worth to be pursued further and expanded in the future.[2] Statistical culture also means having the tools to build trust in official statistics: the trust we must build is based on the clear recognition of the limits of instruments and, at the same time, on the ability to enhance their reliability[3]. Knowledge is a treasure, but studies to extend the quality framework and the capacity to read the data, given the quality framework, are the keys to it.

The answers to the many challenges facing data production at different levels must often come from the **research** field. In the case of official statistics, there is a tension between independence (in the choice of the fields of research) and relevance (lead by the needs of users, producers and existing processes). “There is a need for an effective research program which, on the one hand, has a degree of independence needed by any research program, but which, on the other hand, is sufficiently connected so that its work is both motivated by and feeds back into the daily work of the statistical office” (Fellegi, 2010). Statistical research must lead to improvements in processes and products according to users’ needs with collaboration for experimentation and the development of new techniques . Thematic areas for research will have first of all to be guided by the information needs emerging from SDGs. As stated by the Lamy report (EC, 2017) “The UN Sustainable Development Goals should serve as a global reference framework for defining Europe’s R&I missions”. Yet this must be flanked by the so-called blue skies research. These are the areas of research that apparently have no immediate implications in the real world, but which have anticipated important scientific discoveries more than once in the course of history. Impartiality, objectivity, and scientific independence are the words that must represent the common reference system so that the research activity is carried out in a professional and autonomous way.

Coordination projects should keep on having a pivotal role to better coordinate research activities and avoid duplication of efforts.

Finally, there is a growing need for multidisciplinary research efforts and projects, where subject matter experts (e.g. from the fields of economics, social sciences, etc.) work hand in hand with methodology/statistics experts to find innovative and reliable solutions to important research questions. A current example of such a research effort is the network of distributed, but

[2] See https://ec.europa.eu/eurostat/cros/content/emos_en for details.

[3] By their very nature, sampled and incomplete data might not be considered to be trustworthy by a lay person at first. Awareness has to be raised that such data have been collected and presented with the utmost diligence. Additionally , the fact that certain conditions and budget restrictions simply typically prevent us from taking censuses and that we therefore have to use estimates etc. has to be conveyed to the public. This will build trust.

integrating, research infrastructures called InGRID (Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy), which is currently in its second phase (InGRID-2) and is a collaboration of social scientists and statisticians.[4] A focus on research projects that "only" cover either subject matter experts or methodology/statistics experts is bound to lead to less than optimal results. As stressed by the ESFRI Roadmaps research infrastructures are essential for the promotion of top-level research in all fields. This is particularly true in the field of statistics when the bulk of data to be analysed becomes so huge and complex as in today's world and e-infrastructures, such as those set up by the PRACE[5] project, represent a necessary condition for collaborative research.

[4] http://www.inclusivegrowth.eu/.

[5] http://www.prace-ri.eu.

2. Objectives

Moon-shot objectives:

- Quality and timely data for the full implementation of the 2030 Agenda for Sustainable Development.
- Full comparability of indicators between regions, areas, and also over time for better monitoring and fostering development.
- Availability of evidence-based policy tools at different territorial levels.
- Extension of open data platforms to the whole public administration.
- Move towards a data economy.
- Not only production and dissemination, but statistics as a service.
- Renewed data access resulting from the integration of multiple data sources and e-infrastructures.

Intermediary objectives:

- Exploring new data sources for SDG indicators not measured yet.
- Methodologies for big data treatment and production of related new indicators.
- Methodologies to use big data for the production of indicators at regional level, including small area estimates.
- Coordinating traditional and new data sources on populations not easily represented by the new data but target for SDGs, such as the poor or the victims of disasters.
- Shared big data quality framework.
- Enhanced integration between surveys, administrative data and new sources.
- Develop extended policy tools for social and environmental issues, including macroeconomic models and now-casting techniques.
- Extension of statistical literacy through formal and informal education.
- Improved communication techniques for an increased added value of data.
- Spreading of “smart stats” experiences.

Continuous improvements:

- Increase trust in official statistics.
- Minimise statistical burden of respondents.
- Increase efficiency in data production and dissemination.

3. Themes of action

3.1. New data

Agenda 2030 is an irrevocable strategy which still needs its full set of information for being properly implemented. Eurostat is currently monitoring 100 SDG indicators, yet regular reviews are foreseen as methodologies, technologies and data sources evolve over time (Eurostat, 2017c). Beyond GDP frameworks are spreading in most countries, yet they often still lack the necessary wealth of information. Information on the digital economy, on the global value chains, and on redistribution should be considered as priorities for the near future. Intangible assets are nowadays unavoidable phenomena to be investigated. Despite the huge amount of data already available, our rapidly evolving societies are eager for new data.

FP9 should call for:

- coverage of all SDGs targets. Innovative processes must be introduced for making such an enlarged production cost effective (for example through the use of innovative sources). Research is needed to propose feasible solutions for tier 2 and tier 3 indicators[6];
- better statistics for the globalized world. Social and economic phenomena which are not easily captured with traditional statistical strategies are emerging (such as migration, abandoned minors, disasters, or new forms of organization of economic production through global and multinational value chains);
- timely social and environmental statistics, so that beyond GDP frameworks can be proposed in a continuous way, similarly to the current standard economic indicators;
- the extension of national accounts to social and environmental issues;
- a higher "resolution" of data used as the basis for evidence-based policy and its evaluation. Focussed policy efforts, like the European cohesion policy, need information that is more detailed than state-level data.

[6] Tier 2: The indicator is conceptually clear, has an internationally established methodology and standards are available, but data are not regularly produced by countries. Tier 3: No internationally established methodology or standards are yet available for the indicator, but methodology/standards are being (or will be) developed.

3.2. Methodologies for new sources

Traditional surveys are rather costly and time-intensive and we are facing a fall in the quality of direct investigations (e.g. growth of non-response, decrease in landlines). Nevertheless, no other instrument offers the possibility to gather such detailed and complex socio-economic data on individuals and households. Surveys will have to be complemented by administrative data and other, i.e. new, data sources like big data. In general, statistics is moving towards the utilisation of all available sources, setting enormous challenges in terms of quality and methodologies. In this field, the spaces for future research appear huge with relevant implications not only for official statistics or the measurement of well-being and SDGs, but for the economic system and for entrepreneurial behaviours too.

Innovative methodologies will have to maintain or improve quality, lower the statistical burden and increase the efficiency of all processes.

FP9 should call for:

- a promotion of research within the very diversified field of big data, either being (UNECE, 2013):
- Traditional business systems (process-mediated data, such as scanner data),
- Internet of things (machine-generated data, such as remote sensor data), or
- Human-sourced information (such as social media and web scraping).
- FP9 should foster investigation over availability of sources, information potential and relevance, risks associated with their usage, data quality, and methodologies for estimates production. Improvements must be reached on inference techniques with linked data from multiple data sources;
- evaluation of quality issues relative to the production of statistics based on big data;
- the sharing of experiences and the production of common practices for the use and treatment of administrative data;
- “all-data evolution”: the integration of direct investigations and other sources for a more efficient use of available data;
- further research on statistical data matching methods to support integration processes;
- the production of experimental statistics using new forms of data and a comparison to current systems for producing official statistics. This can also contribute to the debate

about trust in official statistics. Specific applications can include, for example, the estimation of indicators or price statistics.

- an advancement towards a quality framework for big data and new sources, able to fully integrate experimental statistics standards too;
- an increased use of “smart stats” in the whole process of statistical production.
- statistical methods to produce statistics that are comparable over time and between regions. Adding new data sources in the production of statistics or changing the production process of statistical data often has a systematic effect on the outcomes. Methods to avoid confounding real developments and systematic differences due to the introduction of new methods or data sources are important to maintain comparability of outcomes over time.

3.3. Assessment capacity

Statistics is bound to have an increasing role in guiding and evaluating policy action, this being its original role as the “science of the state”. The development of innovative policy tools is fundamental to extending evidence-based decision making through every sector of action. Policy makers do not only need information, they need information on the implications of its use, as well as tailored data. Statistics should somehow be service-oriented.

FP9 should call for:

- innovative policy evaluation tools that are able to fully exploit existing information and that are easily accessible to administrators and stakeholders;
- the introduction of the theme of data integration in statistical modelling;
- extended macroeconomic models able to include social and environmental variables beyond the usual economic ones. To this aim, it is crucial to develop methodologies for the integration of micro and macro sources within models;
- now-casting methods that can be applied to well-being and sustainable development variables, so as to make this information as timely as traditional economic information;
- information at local level, which is rather scarce but fundamental. Methodologies for small area estimation have to be extended to well-being indicators for a better guidance of local policy making.

3.4. Skills and competences development

Widespread knowledge about statistics is a key factor to add value to the produced information. On the other hand, citizens need to have some basic knowledge on the theme to make use of the produced information, recognize quality data and be able to be soundly informed in a world of data deluge and spreading fake news.

FP9 should call for:

- a promotion of methods and tools for enhancing statistical literacy and numeracy through formal and informal education, and a sharing of best practices for teaching statistics and data science at all educational levels. Public campaigns for citizens at large can be realized through new and traditional media;
- Effective communication tools for maximizing data impact and, in particular, SDGs indicators' added value.

3.5. Building a data-friendly environment

Managing and fruitfully utilising the data deluge implies finding an effective regulatory framework for data production, sharing and dissemination.

FP9 should call for:

- open data to become a minimum standard for private and public institutions. Data must be opened further and further to satisfy research needs and to foster transparency;
- simplified, improved and widened possibilities of data access for researchers on the side of data producers. The standard access rights to microdata to a wider audience should allow researchers to be more easily able to work on the microdata of the system, even within the boundaries set by the privacy legislation. In this context, it will be necessary to reach a healthy balance between the legitimate requirements of privacy protection and cognitive potentials of great utility for the public knowledge;
- facilitate e-infrastructures and horizontal data services which would allow for cross-disciplinary usage of information;
- Move towards a data economy. Within the framework of the EU digital agenda FP9 is called to find ways to improve the standards for the use of commercially held data for official statistics, finding incentives for businesses to share data, defining solutions for

reliable identification, exchange of, and differentiated access to data. This implies analysing models of contracts, legislative approaches, privacy standards and modes of use of anonymized data.

References

Data Revolution Group, 2014, *A World that Counts. Mobilising the Data Revolution for Sustainable Development*, Report prepared at the request of the United Nations Secretary-General, by the Independent Expert Advisory Group on a Data Revolution for Sustainable Development. November 2014.

EC, 2016, *Next steps for a sustainable European future, European action for sustainability {SWD(2016) 390 final},* COM(2016) 739 final.

EC, 2017, *LAB – FAB – APP — Investing in the European future we want*, Report of the independent High Level Group on maximising the impact of EU Research & Innovation Programmes. Doi 10.2777/477357.

ECOSOC, 2016, *Transformative agenda for official statistics*, Report of the Secretary-General, Statistical Commission, 16 December 2015, E/CN.3/2016/4

ESSC, 2017, *European Statistics Code of Practice for the National and Community Statistical Authorities,* Adopted by the European Statistical System Committee 16th November 2017. Eurostat and European Statistical System

Eurostat, 2012, Analysis of the future research needs for Official Statistics, *Eurostat methodologies and working papers*. Eurostat

Eurostat, 2014, Toward an harmonised methodology for statistical indicators – Part 1 Indicators typologies and terminologies. *Eurostat Manuals and Guidelines*. Eurostat.

Eurostat, 2016, *Smart statistics*, Task Force Big Data, draft document retrieved from https://ec.europa.eu/eurostat/cros/content/item-4-smart-statistics_en

Eurostat, 2017a, Toward an harmonised methodology for statistical indicators – Part 2 Communication through indicators. *Eurostat Manuals and Guidelines*. Eurostat

Eurostat, 2017b, Toward an harmonised methodology for statistical indicators – Part 3 Relevance of indicators for policy making. *Eurostat Manuals and Guidelines.* Eurostat

Eurostat, 2017c, *Sustainable development in the European Union. Monitoring report on progress towards the SDGs in an EU context*, European Union.

Fellegi, I. P., 2010, The organisation of statistical methodology and methodological research in national statistical offices, *Survey Methodology*, 36, pp. 123-130.

UNECE, 2014, *A Suggested Framework for the Quality of Big Data. Deliverables of the UNECE Big Data Quality Task Team*. UNECE.

UNECE, 2013, *What Does "Big Data" Mean For Official Statistics?*, UNECE.

UN, 2015, *Transforming our world: the 2030 Agenda for Sustainable Development*, United Nations.