Offcial Statistics in the Data Science Worldview: Actors and Stakeholders

Rita Lima[1]
Istituto Nazionale di Statistica
Dipartimento per la raccolta dati e lo sviluppo di metodi e tecnologie per
la produzione e diffusione dell'informazione statistica
Direzione centrale per la metodologia e il disegno dei processi statistici
Servizio Metodi Qualità e metadati
e-mail: lima@istat.it

Abstract

Althought Big Data (BD) provides new methods and ideas for knowledge communities, becoming agents of economic change and of innovation, there is a problem of a new dimension of digital devise between the people who "haves" and "have-nots" skills and competencies to gain new knowledge from handling BD which affects National Statistical Institutes (NSIs). In details, as the private sector collect much of the BD, NSIs have by no means a monopoly on producing statistics while who are in a position of accessing and analyzing data and have the processing power, have the "scientific evidence" to guide environmental policy development. For this reason, nowdays NSIs are under pressure; to stay relevant they have more attention paid to the computer science skills of their statisticians training their staff in this fast-developing field for different aspects of BD related work. In an effort to accommodate these widespread needs, the collaboration of NSIs with the private sector, academia, and civil society could be the way of working to determine the right mix of statisticians and any of the new categories of jobs (i.e. data scientist , DSst). In fact statisticians have proven to be very good in developing and maintaining their data quality assurance frameworks, while relevant data need to be delivered frequently, in a timely manner, and with much detail as DSsts could be available to do. So, is Dsts the next generation actors into this data-driven statistical systems? The answer is not. NSIs should try to keep them in the spotlight with the trusted smart statistics and experimental statistics focusing on these open-ended research topics: a) the border between official statistics, experimental and trusted smart statistics (in not yet easy to feel); b) the external data sources (not always "open"); c) the BD quality (in a basic processing framework). In this fast-moving data landscape, NSIs could play a central role both in making more trust information with expanding the traditional role of statisticians in emerging fields, like DSst. In this respect, the adoption of a data stewardship (DSship) approach enables NSIs to evolve from being a statistics producer to becoming a public service provider that facilitates a joined-up approach to trust in statistical products and statistics across different data and statistics communities.

Introduction

The revolution of data-driven labour market with increasing digitisation of information and the emergence of Big Data (BD) provides new opportunities for employability for young people as *data scientist* (Bušelić and Zorica, 2018; Bejaković and Mrnjavac, 2020). The concept of data *scientist* (DSst), coined in 2008 by D.J. Patil, & J. Hammerbacher, refers to "*people who used both data and science to create something new*". In details the process of learning from a different subject-matter (BD) suggests an emerging intelligence fostered by human interaction with Information Technology (IT) and Digital technologies changing the practice of science and the skilled workforce that must have the following competences: computer science, software development, statistics, data visualization, machine learning, as same as computational infrastructures. However only recently, beyond the availability of massive BD and the intention to utilize them, the needs to be capacity to understand and use data effectively by organizations outside statistical systems are accompanied by

[1] The opinions expressed in this document are the sole responsibility of the author and do not necessarily represent the official position of the Italian National Institute of Statistics.

the difficulties to ensure the truth of data obtained from computer systems for policy or/and development programmes.

Pratically it is not so easy to think about DSst given that in many countries the educational institutions and universities do not even know what the means for career development of data science. So the rich web ecosystem of unstructured data source and BD infrastructures may be not ready to be meet by those who need to learn this growing field.

Moreover althought BD provides new methods and ideas for knowledge communities, becoming agents of economic change and of innovation, there is the problem of a *new dimension of digital devise* between the people who "haves" and "have-nots" skills and competencies to gain new knowledge from handling BD which affects also National Statistical Institutes (NSIs). In details, as the private sector collect much of the BD, NSIs have by no means a monopoly on producing statistics while who are in a position of accessing and analyzing data and have the processing power, have the "scientific evidence" to guide environmental policy development. For this reason, nowdays NSIs are under pressure; to stay relevant they have more attention paid to the computer science skills of their statisticians training their staff in this fast-developing field for different aspects of BD related work: 1.handling meaningful information from BD, 2.making progress on methods, tools, and applications of the various BD sources for official statistics and 3.responding evolved users' expectations on data (easily accessible, available faster, granularity, ect). In an effort to accommodate these widespread needs, the collaboration of NSIs with the private sector, academia, and civil society could be the way of working to determine the right mix of statisticians and any of the new categories of jobs (i.e. DSst). In fact statisticians have proven to be very good in developing and maintaining their data quality assurance frameworks, while relevant data need to be delivered frequently, in a timely manner, and with much detail as DSsts could be available to do. In this fast-moving data landscape, NSIs could play a central role both in making more trust information with expanding the traditional role of statisticians in emerging fields, like DSst.

Challenging time of NSIs in the quest for good data

Meanwhile, as BD are increasingly utilized and integrated into national data and statistical systems (i.e. the experimental statistics), NSIs are trying to keep them in the spotlight with the trusted smart statistics and experimental statistics although there are many open research questions:

- The border between official statistics, experimental and trusted smart statistics in not yet easy to feel;
- External sources are not always «open»
- Continue activities and ongoing projects aim at overcoming criticalities and improving BD quality[2].

In this new data culture, NSIs are also facing several challenges such as: 1. Ensuring data quality, in term of validity and accuracy of their outputs, 2. Building trust and embedding auditable and transparent data life-cycles; 3. Providing adequate anonymization in respect of data subjects' privacy; 4. Creating data partnerships with BD originators; 5. Monitoring and planning for legislation's changes, in compliance with polices, directives and regulations. A quality assurance framework is

[2] In ISTAT, for example, concerning big data and Trusted Smart Statistics it is planned to establish an ad-hoc quality framework. A review of the proposal made in international projects has been made and it will be the basis for developing the Istat proposal (R. Lima, "Dai Big Data alle Trusted Smart Statistics. Nota Tecnica sulla Qualità', Internal Document of ISTAT,luglio 2023).

one of a number of other frameworks, policies and strategies that typically are in place in NSIs to achieve these challenges[3].

In this respect, there is a consensus that the adoption of a data stewardship (DSship) approach enables NSIs to evolve from being a statistics producer to becoming a public service provider that facilitates a joined-up approach to trust in statistical products and statistics across different data and statistics communities (UNECE, 2022). The key assets to be taken into account by NSIs as DSship is "*a set of skills to ensure data are properly managed, shared and preserved, both throughout the research lifecycle and for long-term preservation*". Really helpful to know what DSship is for NSIs, a lot of work that has been done already by them into their everyday practices; it concerns with fiduciary (legal or ethical trust) level of responsibility toward the data, investigating the use of new data sources - including administrative data and BD - and building data science capability for public good (New Zealand Government, 2020; Plotkin, 2021). DSship goes, therefore, beyond the current role of a chief statistician, as it includes giving instructions to all kinds of government and nongovernment institutes regarding the handling of their data. DSship requires a "superset" of analysis capabilities that would be associated with four outcome 'pillars'; governance, collaboration, methods and access which in turn are supported by underlying outputs, activities and inputs/ needs (StatCan, 2021). In summary, all NSIs that are taking care of their own data management according to best principles (the FAIR data principles; Wilkinson, 2019); constitute a large part of the DSship, renewing their role as governmental DSst.

Conclusion

So is DS the next generation ability in a data-driven labour market? The answer is not.

The demand for DSsts as ambivalent people with a curriculum that is mainly build upon a omnia comprensive educational fields (mathematics, statistics, computer sciences, but also sociology, ect.) will definitely narrow, because those people don't exist if not as a team of persons, which sometimes you already have in-house. As discussed by Karen et al. (2020) it needs to include in an interdisciplinary approach the more common technical skills for a DSsts must have (data visualization, programming/software, statistics/mathematics). But many doubts hinge upon that the "soft" skills (communication, data-driven problem solving, critical thinking, learning to give constructive feedback, transparency amongst others,ect.), capabilities and attitudes of DSsts could be necessary in future workers are difficult to teach.

Outside of the NSIs mandate, DSship could potentially be facilitated by a different type of framework of knowledge and skills to use BD, accompanied by a new scope of teaching and learning, a new mode of expertise and all aspects of one's personal and professional lives. For this reason the expanded mandate and function of NSIs to be the DSship could convince us that probably any teaching scheme coud instill such sophisticated traits in a single worker, althought efforts will provide to training. On the contrary, efforts are needed on other factors (e.g. institutions, infrastructure, political stability), for the implementation of policies which could resolve the problems associated with the digital divide in the era of accumulated, available, and accessible BD (Lima, 2017). Such an arrangement creates the opportunity for methodological learning to take place, manage model of social and economic change (i.e. tecnological revolution due to BD and the ageing of population),

[3] The formulation of a quality assurance framework requires an indepth and thorough review of those mechanisms most directly related to quality since the framework's main focus is on the management of the core statistical functions. Statistical laws, regulations and acts, codes of practice, and statistical standards, policies and strategies will need to be explicitly considered, referenced and made readily available in the process of drawing up a quality framework. Frameworks related to quality of statistics had been reviewed mainly in EU countries since the 1990s and created in some countries around 2000. Also, international organizations such as the OECD and IMF established similar frameworks. They listed criterions aimed at statistics including relevance, accuracy, timeliness and specified their checkpoints, sometimes with explanation of the evaluation process, used as a guidance for evaluations.

allowing for a successful reorganization of positive network, more technologically oriented, particularly involving innovation and sharing information and data to overcame high socio-economic disparities between Member States and between regions such as: - dimensions of entreprises (SMEs vs. LE); - historical presence of clusters with diffused tacit unstructured knowledge; - infrastructure and cultural diversity (model of business, approaches, practices, …)
This may not require fashioning new tools and creating neologisms, but adopting a new view sustained by rooted theories and affirmed examples. It seems clear that building knowledge organisational networks, social ("bridging") capital and trust are the new dimensions of knowledge for territorialized spillovers at diverse perspectives: community, neighborhood, firm (Lima, 2017, 2018).

References

Ben-Zvi, D. (2017). Big Data inquiry: Thinking with data. In R. Ferguson et al. (Eds.), Innovating pedagogy 2017. Exploring new forms of teaching, learning and assessment, to guide educators and policy makers (pp. 32–36). Milton Keynes, UK: The Open University.

Bejaković, P. and Mrnjavac, Ž. (2020), "The importance of digital literacy on the labour market", Employee Relations, Vol. 42 No. 4, pp. 921-932.

Bušelić V. and Zorica M.B. (2018). Information Literacy Quest: In Search of Graduate Employability. *Information Literacy in the Workplace*, eds. Serap Kurbanoğlu et al. (Cham, Switzerland: Springer International Publishing, pp: 98–108.

Lima R. (2017). Human Capital for Economic Growth: Moving Toward Big Data. In AA.VV. REGIONAL DEVELOPMENT TRAJECTORIES BEYOND THE CRISIS Percorsi di sviluppo regionale oltre la crisi, Ed. Franco Angeli.

Lima R. (2018). The Big Data Age and the Economies on the Rise: Assessing the Valueof Human Capital in a Data-Driven World, Th thirteenth edition of the Italina National Conference on Statistics, 4-6 July, Rome.

New Zealand Government. (2020). Data stewardship: managing New Zealand's data better to change lives. A Data Stewardship Framework for NZ. Retrieved from: https://www.data.govt.nz/assets/Data-stewardship-framework-and-toolkit-Nov2020.pdf

Plotkin, D. (2021). Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance (2nd Ed.). London, UK: Academic Press.

Statistics Canada. (2021). Statistics Canada's Approach to Data Stewardship [PDF]. Unpublished internal departmental document.

Wilkinson, M. D. (2019) Evaluating FAIR maturity through a scalable, automated, community-governed framework. bioRxiv, doi: 10.1101/649202.

UNECE, 2022: Task Force on data stewardship. Link: https://unece.org/statistics/task-force-data-stewardship