**A knowledge-based approach to the statistical production process**

Stefano Menghinello, Italian National Institute for statistics (ISTAT), Rome, Italy, [menghine@istat.it](mailto:menghine@istat.it)

Claudio Ceccarelli, ISTAT, Rome, Italy, [clceccar@istat.it](mailto:clceccar@istat.it)

Massimo Fedeli, ISTAT, Rome, Italy, [massimo.fedeli@stat.it](mailto:massimo.fedeli@stat.it)

**Abstract**

*This paper explores the nature and key features of the statistical production process in the light of knowledge economics and knowledge management literatures. The business survey is identified as a knowledge intensive process where methodological, thematic, data collection and IT skills and competencies are exploited to improve data quality along different dimensions. The current framework to assess and manage data quality of official statistics seems to pay a limited attention to the knowledge intensive factors of the statistical production process, thus not allowing to fully exploit the information potential of the data in terms of both accuracy and relevance. The knowledge-based elements of the statistical process are introduced in this paper and their implications for process design, data collection and data processing are highlighted to pave the way to a more comprehensive approach to data quality.*

**Keywords:** Statistical production process, knowledge sources and management, data quality.

**1. The role of knowledge in the statistical production process**

In a statistical production process[[1]](#footnote-1), input factors are raw data (from different information sources such as direct reporting, administrative data, big data, etc), the physical capital (mostly IT software and hardware related to data collection, processing and storage), and labour (contributions provided by official statisticians at different stages of the process based upon their scientific and technical skills). The output of a statistical process carried out by a National Statistical Organisation (NSO) is essentially new data labelled as official statistics and made available to data users in different forms, such as basic data, indicators, graphics and maps[[2]](#footnote-2).

A statistical production process is traditionally conceived as a highly standardised process where the value added provided by official statisticians with respect to raw data is bounded to the codification, classification, measurement and statistical representativeness (efficient and unbiased estimates) of final figures, with little impact in terms of new knowledge added to the data.

A wide range of scientific contributions across different academic streams and domains of the economics, innovation, regional, business and management literatures, have emphasised the central role of knowledge factors as key drivers of the quality and innovation of the output for any type of production process. Therefore, these streams of literature can contribute to further develop this discussion within the framework of the statistical production process and data quality assessment and management.

According to these contributions, a production process has both a physical and an immaterial nature, and both should be considered to create the maximum possible value added. As far as the statistical production process is concerned, the relevant body of knowledge includes technological, thematic, data collection and methodological capabilities and skills of the human resources that are only partially incorporated into the physical capital used in the production process, as well as related managerial competencies to set up and maintain efficient and sustainable the production process.

Technological knowledge reflects the capability to design and implement IT infrastructures that support statistical production processes using up to date technologies but also the aptitude to customise standardised IT procedures to the specific statistical needs in terms of process and output requirements. Methodological knowledge encompasses a wide range of scientific methods, standardised tools and more recently conceptual design of production infrastructures, that can support and continuously improve the achievement of statistical goals in terms of data quality.

Thematic knowledge is more difficult to be defined, since it focuses upon different factors: the scientific knowledge on the nature and determinants of the specific phenomena under investigation, practitioners’ skills and routines, including learning process generated by intensive interaction with both data users and survey respondent units. This kind of knowledge is essential at different stages of the statistical production process, such for example the design of the questionnaire, the definition of data collection strategies, the selection of the most appropriate parameters for methodologies and practical solutions to be applied in day by day production routines[[3]](#footnote-3).

Data collection knowledge encompasses all techniques, methods and organisational procedures that can make large scale data collection more efficient and effective, also reducing the burden on the respondents. It also includes the capability to customise different models of data collection by interacting with thematic experts.

Two contributions from the knowledge economics and management literatures need to be recalled here to further develop the discussion on the nature and the management of the statistical production process as a knowledge intensive one.

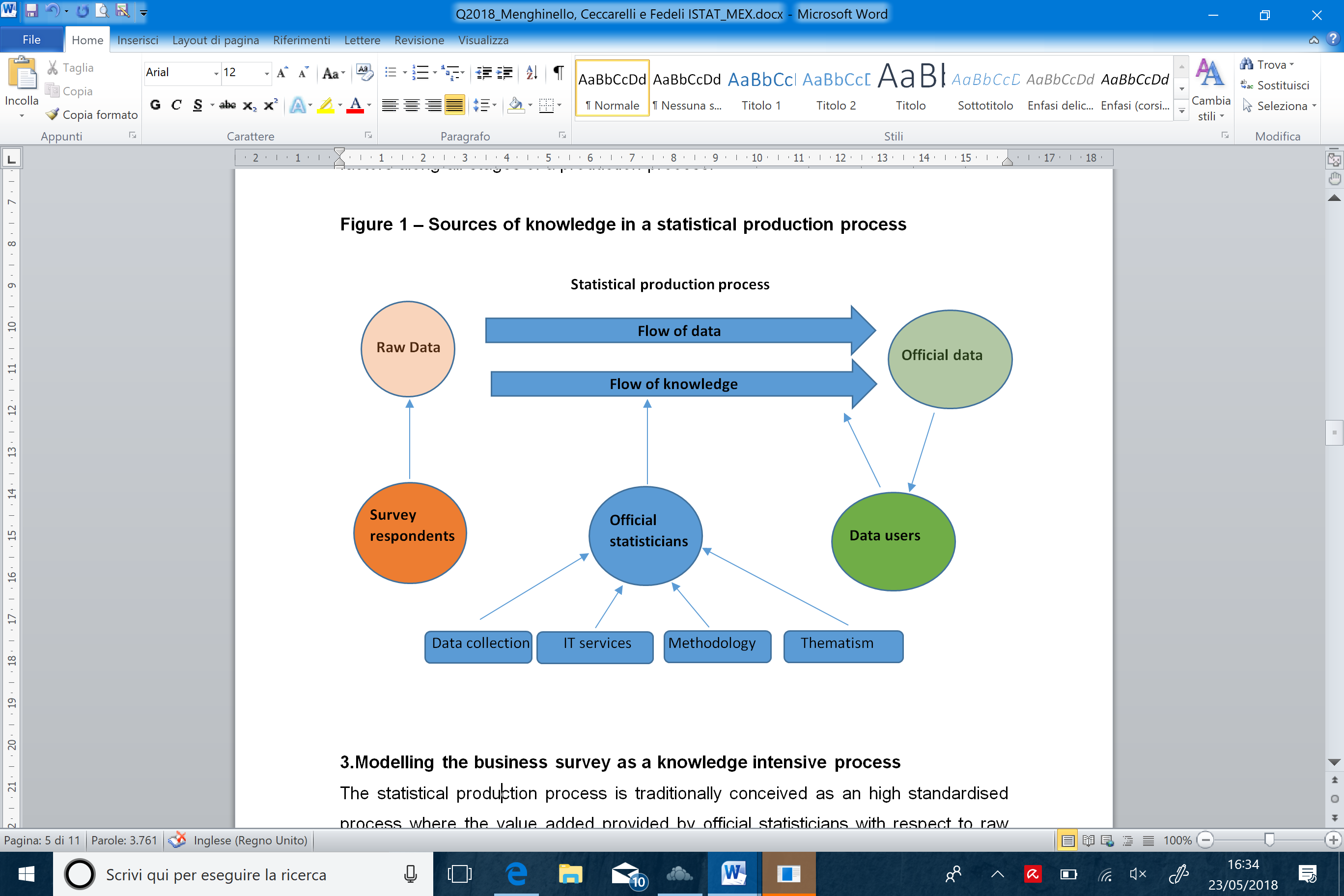
The term "tacit knowledge", first attributed to Michael Polanyi (1958) and then further developed by other scientists, such for instance Nonaka (1987, 1991), encompasses skills, ideas and experiences that people actively engaged in a production process do effectively use both on daily routines as well as to introduce innovations, but that are not codified and may not be easily transferred. With tacit knowledge, people are not often aware of the knowledge they possess or how it can be valuable to others[[4]](#footnote-4).

Given the presence of relevant differences in the design, organisation and maintenance of statistical production processes across countries and quite often within the same NSI, it is self-evident that only a superficial view can exclude the relevance of tacit knowledge in statistical production processes, since both day by day actions to ensure data quality and medium-long terms actions finalised to improve data quality heavily rely on both explicit and tacit knowledge[[5]](#footnote-5).

Knowledge management typically focuses upon the management of knowledge as a strategic asset of a given organisation[[6]](#footnote-6). It has a multidisciplinary nature where organisational, technological, and social dimensions are jointly considered. Key issues considered by this multidisciplinary approach include the definition of the most appropriate strategic and operational guidelines to enhance the transfer of tacit knowledge, the identification and development of different types of learning processes, and the design and implement an organisational framework that can efficiently support the strengthen of knowledge factors along all stages of a production process.

The implications of this approach when applied to the statistical production process are quite straightforward. A statistical production process involves complex interactions between technological, methodological and data processing routines designed and managed by a community of official statisticians with different skills and competencies that are only partially codified as explicit knowledge. In addition, learning processes activated by the interaction with data users (user’s needs in terms of accuracy, relevance and new indicators) and data providers (feedbacks from survey respondents in terms of feasibility, reliability and accuracy of requested data) need to be properly codified in order to incorporate this knowledge into the statistical production process. Therefore, different sources (internal versus external) and types of knowledge that are potentially valuable to improve the quality of output of a statistical production process need to be properly identified and managed (Figure 1).

**Figure 1 – Sources of knowledge in the statistical production process**



**3.Modelling the business survey as a knowledge intensive process**

The distinction between a knowledge transfer and a knowledge creation process is of foremost relevance when one addresses the issue of modelling the business survey as a knowledge intensive process. In the first case, the knowledge embedded in data collected from respondent units is shifted to final users with no alteration in data processing (1), while in the second case the knowledge is transformed by the process to generate more consistent, accurate and business relevant data using other inputs, such as labour, physical capital and technology (2).

(1)

(2)

Where Q represents the official output of a statistical process, D= raw data, L=labour, K= physical capital and A=technology. The parameter is a simple transformation parameter that converts raw data into official statistics with no alteration in the informative power of the data connected to the exploitation of additional sources of knowledge.

The contributions introduced in the previous sections, based upon the knowledge economy and knowledge management literatures tend to acknowledge that any production process is knowledge based, since the knowledge embedded in raw inputs (being goods or services) is either explicitly (codified knowledge) or implicitly (tacit knowledge) transformed into the final output by the human resources, physical capital and technology used in the process.

As a result, equation 2 can be considered the most appropriate approach to model a statistical production process. The Cobb-Douglas is one of the most popular forms of production function used in both theoretical and empirical analysis.

(3)

Where P is the value added generated by L e K inputs in the production process, A is an efficiency coefficient that measures total factor productivity. Although this production function is widely used in theoretical and empirical research, it not suitable to measure the contribution of different types of knowledge, since it is essentially devoted to assess the efficiency level of a production process with respect to which knowledge is essentially an exogenous time invariant variable. A substantial advancement in this respect is provided by the work of Pakes and Griliches (1979,1984) and Griliches (1990) that explicitly try to model a knowledge production function (KPF) based on a Cobb-Douglas-type production function specification:

** (4)

The KPF describes the relation between knowledge inputs on the one hand and knowledge output on the other. Therefore, KI are knowledge inputs and KO is the knowledge output. Within the framework of official statistics, knowledge inputs can be identified as the skills and capabilities of researchers and technicians devoted to the production process related to their different areas of expertise as highlighted in the previous section[[7]](#footnote-7) while the knowledge output is the value added to raw data in terms of new/better knowledge embedded in the data[[8]](#footnote-8).

**4. Some implications of this approach for data quality**

The adoption of a knowledge-based approach paves the way to a more general reconsideration of the key features of a statistical production process and their link to data quality assessment and evaluation.

In this respect, the traditional approach in questionnaire design that tends to impose to the respondent units an highly codified questionnaire, indeed a limited set of predefined alternative options for each question based on standard statistical definitions, is both restrictive and burden-generating. It is restrictive since it implicitly imposes a strong limitation to the amount of knowledge that can be sourced from respondent units in a given span of time with respect to a specific phenomenon. It is also burden-generating, since the adoption of a technical language (set of statistical definitions) finalised to assure comparability and accuracy is not always easily understandable by respondent units, thus generating a knowledge burden (costs to understand, retrieve and reclassify business concepts according to statistical definitions) that can be more intense than traditional burden determinants with uncertain effects on the accuracy of collected data[[9]](#footnote-9).

In a similar vein, both the knowledge collected from respondent units and the knowledge generated by official statisticians tend to remain constrained to production process routines (*tacit knowledge*) and it is usually not disseminated as data or metadata to data users to enrich or at least complement the standard dissemination of official statistics. In conclusion, once the knowledge nature of a statistical production process is fully acknowledged, not only processes but also data collection, data processing and data dissemination strategies shell be reshaped to take into account this new relevant dimension.

**5. GSBPM and the knowledge based approach**

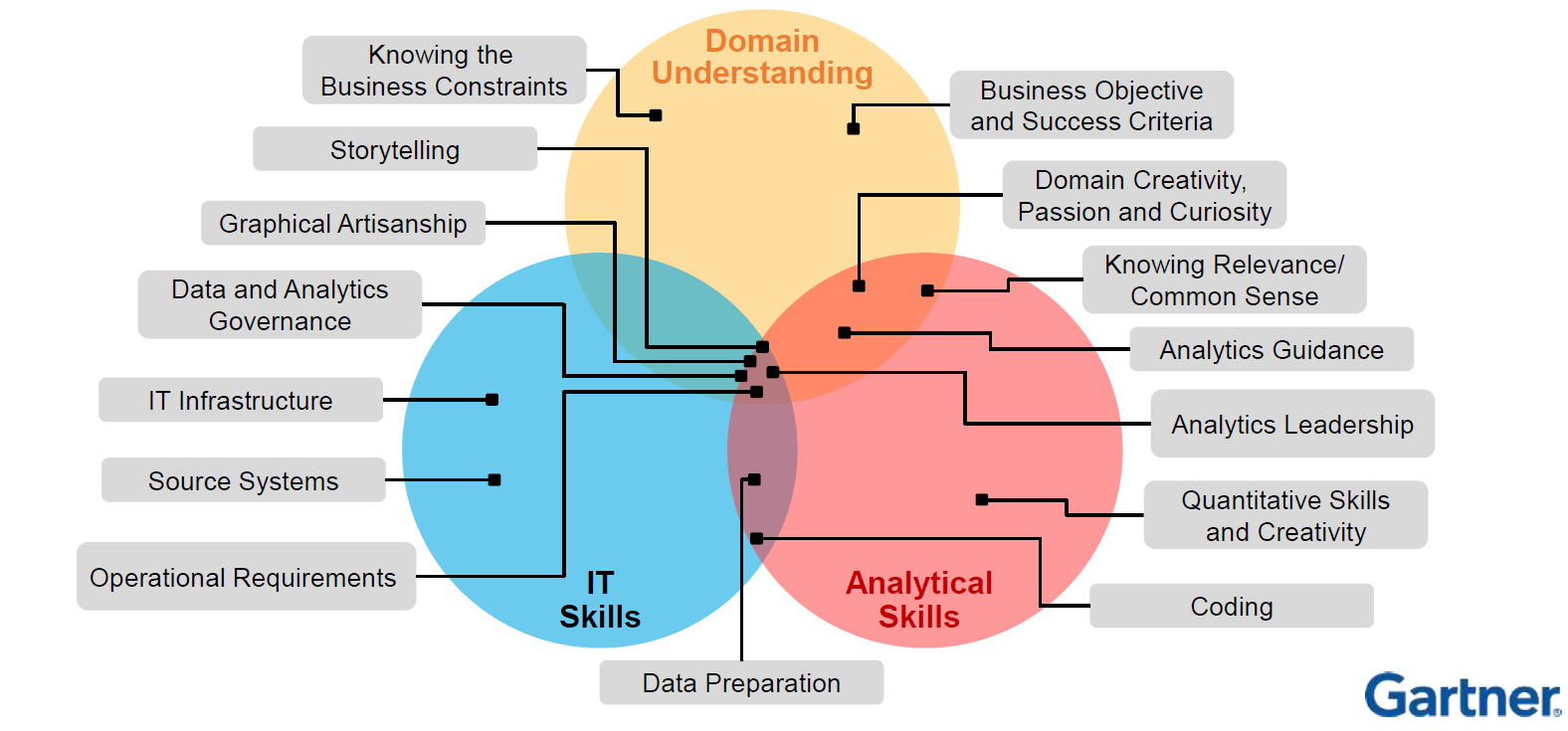
The introduction of the GSBPM framework is of foremost interest for this paper[[10]](#footnote-10). Firstly, it allows to map all different stages of a statistical production process independently from the peculiar characteristics of a specific statistical domain. Secondly, it paves the way to further development of this research work since all relevant sources of internal/external and tacit/explicit knowledge associated to a specific stage or task of a statistical production process could be linked to a GSBPM framework and thus be used to design and implement not only strategic but only operational strategies based upon the strengthening of knowledge-based factors. As an example, the goal to produce more business relevant data and/or to reduce the statistical burden related to knowledge distance between official statisticians and respondent units requires to adopt specific actions to improve the quality of data and to minimise the impact on respondents. These actions will be based upon knowledge upgrading processes related to specific tasks/stages of a statistical production process (questionnaire design, data collection strategies and operations) with a strong interaction across IT, data collection, methodology and thematic expertise and competencies within NSIs.

**6. The role of IT for knowledge intensive statistical production processes**

The set up and maintenance of a well designed and efficient information system traditionally represents the back bone of any reliable and high quality statistical production process. Since its foundation, knowledge management (KM) has been strictly connected to IT science and operations. Indeed, the core components of KM include people/culture, processes/structure and technology. It then become natural to explore all possible interactions between an information system and a statistical production process according to the knowledge-based approach. By large, the knowledge-based approach to statistical production process calls for the design and implementation of IT services that are capable to manage both the physical flows of data and the immaterial flow of knowledge.

We believe that a holistic approach join IT skill, methodological skill and domain skill together to obtain better knowledge management (KM) (Figure 2).

**Figure 2 – Holistic approach for the knowledge management**



An interesting categorization in this respect distinguishes embedded knowledge of a system outside a human individual (e.g., an information system may have knowledge embedded into its design) from embodied knowledge representing a learned capability of a human being. As explained in the second section, a large share of knowledge used by official statisticians in a statistical production process is tacit and tends to be dissipated if not properly managed. Therefore, it cannot be easily codified in IT procedures. Alternative solutions to retain such knowledge although in a rather unstructured way are knowledge repositories (databases, bookmarking engines, etc.)

In addition, learning processes based on knowledge exploitation tend to be dynamic in contrast to statistical codified rules included in standard IT procedures. For, instance, data validation process grounded on standardized methodological techniques are usually not sensible to the thematic knowledge developed by technicians devoted to manual data quality checks.

As a result, learning processes based upon the interaction between standard methodological procedures and specific data validation rules developed by technicians based upon their own knowledge and experience cannot take place unless they are explicitly designed and implemented in an advanced information system. Dynamic data validation rules and the design and implementation of advanced user interfaces that allow statisticians to interact with automatized methodological procedures by incorporating revised parameters based on their data quality manual checks will support learning processes by generating feedbacks and interactions between automatic and thematic manual rules.

**References**

Fritsch, M. and Slavtchev, V. (2006) "Measuring the efficiency of regional innovation systems: an empirical assessment," Freiberg Working Papers N.08, Bergakademie Freiberg, Faculty of Economics and Business Administration.

Girard John P., Girard JoAnn L. (2015). "Defining knowledge management: Toward an applied compendium" (PDF). Online Journal of Applied Knowledge Management. 3 (1): 14.

Griliches Z. (1990). Patent statistics as economic indicators: a survey, Journal of Economic Literature 20(4), 1661.

Griliches Z. (1970). “Notes on the Role of Education in Production Functions and Growth Accounting,” in L. Hansen, ed., Education and Income, NBER Studies in Income and Wealth, Vol. 35, New York, 1970.

Gupta, Jatinder; Sharma, Sushil (2004). Creating Knowledge Based Organizations. Boston: Idea Group Publishing. ISBN 1-59140-163-1.

Logan V.A., Davis M., Sicular S. and others. Domain Analytics: Harnessing the Pervasive Nature of Analytics. Gartner (G00300139).

Maier R. (2007). Knowledge Management Systems: Information And Communication Technologies for Knowledge Management (3rd edition). Berlin: Springer.

Murray E. Jennex (2009). Knowledge Management, Organizational Memory, and Transfer Behavior: Global Approaches and Advancements, Information Science Reference.

Nonaka I. (1987). Managing the ﬁrm as information creation process. Working paper, Institute of Business Research, Hitotsubashi University, Hitotsubashi, Japan.

Nonaka I. (1991). The knowledge-creating company. Harvard Bus. Rev. 69(6) 96–104.

Pakes A., Z. Griliches (1984). Patents and the R&D at the firm level: a first look, in: Z. Griliches (ed.), R&D, patents and productivity, Chicago: Chicago University Press, 390-409.

Polanyi M. (1958). Personal Knowledge: Towards a Post-Critical Philosophy. University of Chicago Press. ISBN 0-226-67288-3.

Sallam R. L., Howson C. and Idoine C. J. Augmented Analytics Is the Future of Data and Analytics. Gartner (G00326012).

Sanchez R (1996). Strategic Learning and Knowledge Management, Wiley, Chichester.Wright Kirby (2005). "Personal knowledge management: supporting individual knowledge worker performance". Knowledge Management Research and Practice. 3 (3): 156–165.

1. A production process can be defined as any process that transforms inputs (goods and/or services) into outputs (goods and/or services) that are valuable for users. Labour and physical capital are traditionally considered the key input factors, in addition to raw materials and intermediate goods and services. [↑](#footnote-ref-1)
2. Extensive technological and organisation innovations - such as web based data collection and more intensive use of administrative data, improved IT capabilities and reshaping of internal organisation according to the BA principles - have led many NSI to substantially improve the efficiency of their statistical production processes. As a result, labour and physical capital factors are not as important as before in terms of quantity, but they have become more relevant in terms of quality, since they can also be explicitly linked to concrete actions finalise to improve the different dimensions of data quality, with a stronger emphasis on accuracy and relevance. [↑](#footnote-ref-2)
3. In the broad domain of business statistics, thematic knowledge can be relevant, for instance, in discriminating the nature of outliers (wrong data versus anomalous data that are anyhow correct since they can be explained by enterprise or product characteristics). This is the case of foreign trade statistics, where thematic knowledge is focused on products and exporting enterprise characteristics. Thematic knowledge it is also important in the validation of data, such as short terms indicators (thematic knowledge on non-seasonal component of firm level variability) and structural business statistics, with a specific emphasis on R&D and FATS statistics (thematic knowledge on the organisation, ownership structure and R&D location of multinational enterprises). Thematic knowledge is playing an increasing role in business statistics (Profiling, upgraded definitions of statistical units, Large Case Units) to improve the accuracy and relevance of official statistics, and it is therefore essential to include this important dimension of data quality in a generalised framework of the statistical production process. [↑](#footnote-ref-3)
4. The key to acquiring tacit knowledge is experience. Without some form of shared experience, it is extremely difficult for people to share each other's thinking processes. Effective transfer of tacit knowledge generally requires extensive personal contact, regular interaction and trust. In contrast, formal, codified or explicit knowledge refers to knowledge that can be readily articulated, codified, accessed and verbalized, and thus it can be easily transmitted to others. Most forms of explicit knowledge can be stored in books, manuals, and operational procedures. [↑](#footnote-ref-4)
5. This approach is not necessarily in contrast with the adoption of a single statistical process approach, since tacit knowledge is not necessarily related with inefficiency, duplication and high costs of management traditionally associated with the high fragmentation of statistical production processes in many countries. [↑](#footnote-ref-5)
6. Knowledge management (KM) is a new discipline that analyses the process of creating, sharing, using and managing the knowledge and information of an organisation (Sanchez, 1996; Gupta and Sharma, 2004; Sushil, 2004; Wright, 2005; Maier, 2007; Jennex,2009; Girard, J. P. And Girard, J.A L., 2015). [↑](#footnote-ref-6)
7. Knowledge source for outside trough interactions with data users and input data providers (survey respondents) can be modelled as knowledge spillovers. Other variables should also be included in the model to take into account the contribution of managerial capabilities to improve the value added of the data. [↑](#footnote-ref-7)
8. The measurement of the new knowledge incorporated in the data seems the most complicated task to be achieved. On one side, given the public nature of official statistics, a direct and monetary quantification of the value of official data is quite difficult to be assessed. On the other side, the standard statistical components of data quality in official statistics are not completely orthogonal with respect to the knowledge-based ones.The use of national comparison with respect to other available data sources that measure similar topics, international comparison (asymmetries studies within the framework of official statistics), and direct reporting on data users and data providers to assess the specific value added produced by NSI (institutional reputation, capability to address complex measurement problems both in terms of data collection and data dissemination, etc) could be quite useful in this respect. [↑](#footnote-ref-8)
9. An alternative approach shell consider the possibility to collect more unstructured but more business relevant data from respondent units that can generate less burden on the respondents and could be synthesised trough appropriate statistical methods with a more careful assessment of the accuracy of data, virtually impossible in the case of highly codified questionnaire where all the uncertainty and little accuracy in the data classification process according to statistical concepts and definitions is fully incorporated in the respondent units cognitive process and thus unobservable by official statisticians. [↑](#footnote-ref-9)
10. The UNECE statistical community, jointly with other international organisations and country best practices, has recently developed a set of statistical standards (CSPA, GSBPM, GSIM, GAMSO) to boost the modernization of core activities of a NSIs: production processes and technical support activities such as data collection, IT and methodology. In parallel to that, the creation of the ESS 2020 Vision and associated management structures to steer the development of European Statistics (ESS 2014) have been introduced. [↑](#footnote-ref-10)