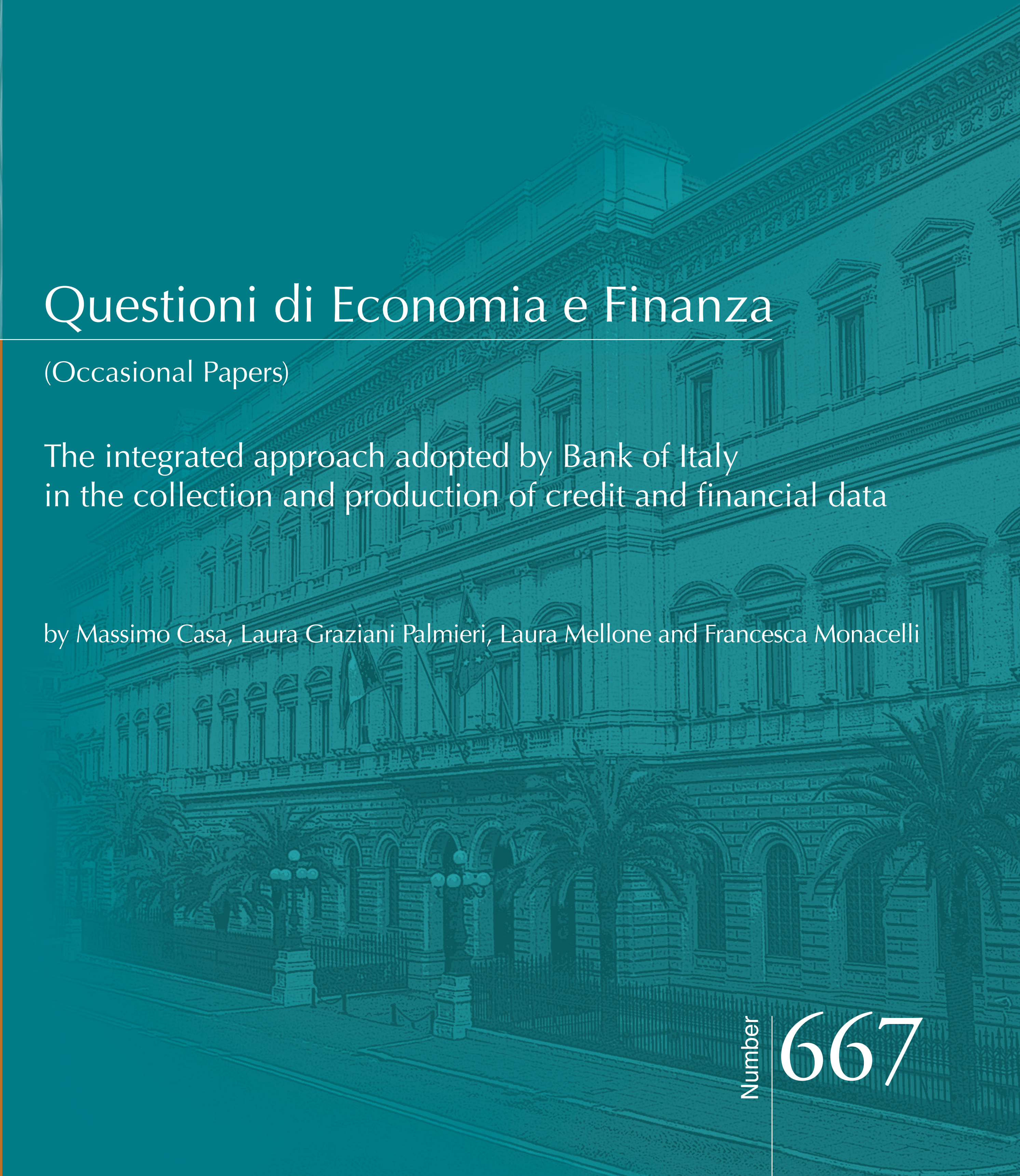
BANCA D'ITALIA
EUROSISTEMA



Questioni di Economia e Finanza

(Occasional Papers)

The integrated approach adopted by Bank of Italy in the collection and production of credit and financial data

by Massimo Casa, Laura Graziani Palmieri, Laura Mellone and Francesca Monacelli

February 2022

Number 667

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

The integrated approach adopted by Bank of Italy in the collection and production of credit and financial data

by Massimo Casa, Laura Graziani Palmieri, Laura Mellone and Francesca Monacelli

Number 667 – February 2022

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at *www.bancaditalia.it* .

ISSN 1972-6627 (print)
ISSN 1972-6643 (online)

Printed by the Printing and Publishing Division of the Bank of Italy

THE INTEGRATED APPROACH ADOPTED BY BANK OF ITALY IN THE COLLECTION AND PRODUCTION OF CREDIT AND FINANCIAL DATA

by Massimo Casa[*] , Laura Graziani Palmieri[*], Laura Mellone[*] and Francesca Monacelli[*]

Abstract

The paper illustrates the phases of the process that the Bank of Italy follows to produce the statistics derived from credit and financial reporting: the identification of the information requirements; the definition of the data model; the design of the new data collection method to be used by reporting agents; cooperation between the Bank of Italy and reporting agents; data quality procedures; and dissemination of this information to internal and external users. This process takes an 'integrated approach' and was adopted by the Bank of Italy in the late 1980s. For the last decade, it has been a reference point for the European System of Central Banks both as regards the development of the statistical framework and for the efficiency improvements in data management and data governance on the part of the authorities. The Bank of Italy takes part in these initiatives providing a valuable contribution in terms of ideas and experience.

JEL Classification: C81, G21, M15.
Keywords: regulatory reporting, banking reporting, data model, data quality, information management, statistical production, information system.
DOI: 10.32057/0.QEF.2022.0667

Contents

[*] Bank of Italy, Statistical Data Collection and Processing Directorate.

Introduction[1]

The Bank of Italy produces many statistics, including credit and financial statistics, derived from the data sent by intermediaries …

The Bank of Italy is institutionally responsible for the production of numerous macroeconomic statistics, in particular regarding: credit and finance; the balance of payments; financial markets; and public finance. Data supplied by intermediaries (i.e. banks and other financial institutions) represent a major information source for such statistics, together with data from the entities and securities registers, the information acquired through the sample surveys on households and on businesses and the data flows received from external sources[2]. Such a wealth of data constitutes the basis on which the Bank founds its analysis and actions.

The role of the Bank of Italy as a producer of statistics based on the data supplied by intermediaries derives from national and EU legislation, which both assign powers to the Bank that complement the role of the National Statistical Institute (ISTAT).

… on the basis of a production process in which it is necessary to balance some aspects

The production process of these statistics is made particularly complex by the need to balance some important aspects that may seem in contrast with each other. We refer, in particular, to: the timeliness of the availability of data and its quality; the costs associated with the production, collection and dissemination of information and its utility; and the search for a higher level of detail of the data (so-called granularity), which increases the administrative burden for the reporting agents and for the Bank of Italy itself.

The complexity of the process has grown in tandem with the increasingly international dimension of statistical production …

Some events have accelerated the cross-country harmonisation of the regulatory reports applied to banks and financial intermediaries. We refer to the participation, since the early 1990s, of National Central Banks (NCBs) in the European System of Central Banks (ESCB), the more recent participation of national supervisory authorities in both the European System of Financial Supervision (ESFS) and the Single Supervisory Mechanism (SSM), and the participation of the national banking resolution authorities in the Single Resolution Mechanism (SRM). This process also led to the search for a balance between national information needs, necessary to undertake domestic tasks, and those arising from European legislation.

… and with the dynamism of the reference scenario

The overall context is highly dynamic. The perimeter of the information requested by national authorities tends to expand along with: (1) the diversification of the phenomena of interest, e.g. of the new statistical needs regarding Fintech, the green economy and sustainable finance; (2) the new types of intermediaries subject to reporting obligations; and (3) the growing availability of non-conventional data complementing official statistics. The overall level of complexity is itself the subject of a heated debate between the NCBs, the European Banking Authority (EBA), the European Central Bank (ECB), the Single Resolution Board (SRB) and the banking industry. The debate aims, among other things, to develop in Europe an integrated system for the collection of statistical, resolution and prudential data from banks and financial intermediaries.

[1] We would like to thank Roberto Sabbatini for his valuable comments. The views and the conclusions are solely those of the authors and do not necessarily reflect the views of the Bank of Italy.

[2] "External sources"refers to the data acquired from commercial providers, other national and international authorities and institutions.

What we have just described suggests how sensitive and complex the data production process is and how important the step preceding the use of the data is, whether carried out by internal or external parties vis-à-vis the Bank of Italy.

The work describes the fundamental stages of the process followed by the Bank of Italy to produce credit and financial statistics derived from reports from intermediaries …

This paper describes the fundamental phases of the process adopted by the Bank of Italy to produce the statistics mainly derived from credit and financial intermediaries' reports. This wealth of data is fed with data reported by a wide range of intermediaries such as banks, money market funds, other deposit-taking corporations, payment services providers, financial institutions, and other non-bank financial intermediaries (e.g. non-monetary mutual funds, asset management companies, credit securitization companies, and leasing, factoring and consumer credit companies). The content and level of detail of the data provided by the above-listed institutions have progressively expanded to meet the evolving needs of their multiple users. In this respect, it is important to specify that the users are not only located in the Bank of Italy, but they also belong to other national (e.g. ISTAT and CONSOB[3]) and international institutions and authorities to which the Bank of Italy provides significant data flows.

… and to deal promptly with new information needs

The method adopted to govern and manage this information asset has, over time, allowed the Bank to respond effectively, and promptly, to new requests for information associated, for example, with the launch of the European Monetary Union (EMU) and the creation of the SSM and the SRM. Moreover, it has helped to fill the important data gaps that emerged after the global financial crisis of 2008-09, as identified in the context of the G20 Data Gaps Initiative[4]. More recently, the availability of granular information has also made it possible, making only a few adjustments, to deal with the new information needs arising from the outbreak of the Covid-19 pandemic. This resulted in quickly being able to provide policy makers with a thorough insight into the Italian economy for them to assess the impacts of the measures adopted in Europe to support credit to households and businesses and to safeguard the stability of the financial system[5].

The integrated approach is a key aspect …

The model adopted by the Bank of Italy since the late 1980s for the production of credit and financial statistics is based on the "integrated approach". This approach stems from the consideration that a single item of data can be exploited for multiple uses and, therefore, it must be requested and collected only once. It is also based on a holistic vision of the corporate statistical information system, according to which the data are governed and managed by common organizational units, methodologies and IT infrastructures, which are at the service of multiple users. This approach is diametrically opposite to the operative separation of the different statistical domains, which replicate surveys, processes and tools – the "silos approach". On the contrary, the integrated approach favours the identification and exploitation of the synergies and relationships between different segments of the Bank's information assets. In so doing it aims to avoid, as far as possible, the duplication of data requests to reporters, thus helping to contain the reporting burden. Moreover, it increases the possibility of

[3] They are respectively the Italian National Statistical Institute and the Italian Securities and Exchange Commission.

[4] In the 2009 survey carried out, at the request of the G20, by the International Monetary Fund (IMF) and the Financial Stability Board (FSB), the areas for improvement and the information gaps identified concerned the statistics to monitor: 1) financial system risks (such as those on financial leverage, liquidity, maturity transformations and on recourse to risk transfer instruments by intermediaries), 2) the interconnections in international financial networks (especially in terms of flows of Systematically Important Financial Institutions (SIFIs), cross-border banking flows and activities of non-bank financial intermediaries); and 3) the vulnerability of households (through sectoral statistics, such as on those of financial accounts, public finance, on the prices of real assets). See in this regard: Financial Stability Board and International Monetary Fund (2009).

[5] Casa M., D'Alessio G. (2020).

the cross-use of different information segments and, from a technological point of view, it also supports the application of significant economies of scale. In brief, the fundamental ingredients of this approach, that allow us to facilitate the joint use of data and maximise its exploitation, are: (1) the multi-purpose use of the same data (e.g. for monetary policy, supervision, financial stability, payment system oversight, and consumer protection); (2) the adoption of unambiguous, common definitions and codes to describe the reporting schemes (i.e. a single statistical dictionary); (3) a corporate statistical data warehouse (DWH); (4) company-wide data governance; and (5) a dedicated IT platform supporting the management and use of the data. At the same time, within the framework of the integrated approach, it is also possible to move away from the fully holistic vision described above and to implement the approach only partially. This helps to limit the time needed to set up new data collections and it is particularly useful when dealing with very urgent information needs or data requests that are deemed as only occasional. Should the data collection become steady and regular, it is then always possible to return to the fully integrated scheme.

The paper is organised as follows. Section 1 provides an overview of the evolution of users' information needs regarding banking and financial data. Sections 2 to 5 analyse the different aspects of the integrated approach. In particular: Section 2 focuses on the unitary and coordinated detection of users' information needs; Section 3 describes the fundamental characteristics of the multidimensional data model and the key role played by the integrated statistical dictionary in guiding the collection, storage and navigation of data in the DWH; Section 4 examines the issues of quality management, also focusing on cooperation initiatives with intermediaries; and Section 5 illustrates the advantages of the single corporate DWH and dissemination policy. Lastly, Section 6 is devoted to a discussion of the future challenges. The Glossary illustrates the main technical terms used in the paper[6].

1. The evolution of the users' information requirements

Until the 1990s information needs had mainly a national dimension

The design of a new data collection, or the enhancement of an existing one, is called "information project" and it is always triggered by the users' need.

Until the 1990s the information requirements have predominantly been the expression of national needs which mainly referred to prudential supervision on banks and financial intermediaries and to economic analysis to support monetary policy decisions. Eventually they have been complemented by the requirements arising from the oversight of payment systems and financial markets tasks, which are also assigned to the Bank of Italy.

Subsequently the European information needs emerged … and not only

Subsequently, the information needs acquired also a European dimension, whose importance grew over time together with the progress of the integration process that resulted in the launch of the single monetary policy and, later on, in the reform of the European architecture of the micro and macro prudential supervision and the establishment of the SSM and SRM. With the data gaps emerged after the global financial crises at the end of the first decade of the current century and the initiatives

[6] Terms contained in the Glossary are underlined the first time they appear in the main text.

of the G20, Financial Stability Board (FSB) and International Monetary Fund (IMF), new requirements for credit and financial data, and the related information projects, have transcended the European dimension to become global.

The need for more granular data

As to the data collections carried out by the Bank of Italy, this process is characterised by the growing demand for increasingly granular data[7] that are deemed necessary to better analyse the heterogeneous behaviors of economic agents and to facilitate a more in-depth assessment of their economical interdependencies. At present, the European and supra-European dimensions of the data needs are predominant and influence the entire design of the statistical collections.

In brief, since the beginning of the current century more and more data collections from intermediaries were progressively added to the original set of information that the Bank of Italy manages on behalf of the ECB (i.e. balance-sheet data and interest rates of monetary and financial institutions and data relating to the shadow banking sector). They include: harmonised supervisory data defined by the EBA (initially these data consisted in capital, liquidity and financial leverage and subsequently they were extended to many other information domains used for banking supervisory and resolution purposes); several reports for the Single Resolution Board[8] and the ECB. Among the latter, the data collection on individual securities holdings by the different institutional sectors and banking groups (so called Securities Holdings Statistics, SHS)[9], the granular survey on loans granted by the European banks (AnaCredit[10]) and the granular survey on Money Market Statistical Reporting (MMSR[11]) stand out.

The list is not complete. The growth of the data reported by banks and financial intermediaries to the Bank of Italy stemming from the European and international side is enormous. In particular, as at June 2021, 70% of the 163 different data collections are in place to meet the needs of supranational authorities. In 2000, the number of data collections was significantly lower (30), of which only 3 were established to provide data to ECB.

The registers take on a leading role

Together with the increase in the granularity of the data collected, the importance of the registers has also grown over time, i.e. the archives of the characteristics of financial products (securities, derivatives, etc.) and of the institutional units (banks, individuals, businesses, etc.). They are, respectively, the Securities register and the Entities register. The registers represent the foundations for the granular data collections (including data collections on individual entities). At the same time, they are the indispensable tool to integrate the collected data, thus extraordinarily enriching their information value[12].

[7] See M. Draghi. (2016), "Welcome address at the Eight ECB Statistics Conference: Central bank statistics: moving beyond the aggregates", Frankfurt, July 6th: *"Disaggregated data are indeed necessary to identify and analyse the heterogeneity that characterises the real world. For central banks this is particularly important: to implement policy in the most effective way, we need to know how our policy actions affect all sectors of the economy. Both the challenges posed by the current economic climate for monetary and macroprudential policy, and the information required to carry out microprudential supervision by the Single Supervisory Mechanism (SSM) increase our need for granular data*".

[8] The Single Resolution Board is a key authority of the Banking Union and in particular of the Single Resolution Mechanism, that comes into action in the event of bankruptcy or risk of bankruptcy of a bank of the Eurozone or of the States that adhere to the Banking Union. It has the task of ensuring the orderly management of bankrupt banks with the lowest possible costs for taxpaying citizens and with minimum impact on the real economy.

[9] See https://www.ecb.europa.eu/stats/financial_markets_and_interest_rates/securities_holdings/html/index.en.html.

[10] See https://www.ecb.europa.eu/stats/money_credit_banking/anacredit/html/index.en.html.

[11] The MMSR surveys provide for the daily reporting of individual transactions carried out by a sample of major banks in the euro area on some money market instruments; since 2019 they are also used for the calculation of the Euro Short Term rate (€STR). See https://www.ecb.europa.eu/stats/financial_markets_and_interest_rates/money_market/html/index.en.html.

[12] With reference to financial products, very analytical data on a security-by-security basis, have been included since 2004 in the Centralised Securities Database (CSDB) of the ECB which, for the Italian securities, is fed by the Bank of Italy with the information available in the

The evolution of structured and periodic reporting by intermediaries

Structured and periodic reports, which banks (since the early 1980s) and other financial intermediaries (since the 1990s) are required to transmit, are now characterised by a significant wealth of details which reflect the innovations introduced over time to satisfy new and differentiated requests from users. The users may be very different from each other and therefore interested in different angles and perspectives of the same piece of information. For example, financial innovation continuously prompts new and legitimate information needs for authorities to carry out their economic analyses and to disseminate data to the public at large. Likewise, the diversification of the distribution channels of banking products has consolidated over time the need to obtain further information in bank supervisors and payment systems controllers.

In brief, the picture will surely further and rapidly evolve. At the present juncture, new information needs are looming on the horizon that will require the introduction of new data collections: climate change, sustainable finance and Fintech data, as well as information on payment services.

2. The coordination of the various information requirements

The approach adopted by the Bank of Italy does not know the "silos"

As mentioned earlier in the introduction, since the late 1980s in the Bank of Italy the users' needs are not satisfied according to a silo logic. The latter, in fact would entail that each information segment is treated separately from the others and that the emphasis is applied to end-uses rather than to the optimization of the *ex ante* production and compilation phases. Instead, the followed approach aims at putting together the information needs of the various institutional functions and then acts in a coordinated way. More in detail, the principle adopted by the Bank of Italy is that, by going in-depth in the granularity level of the data, it is possible to describe a phenomenon in such a way that it can be exploited to satisfy multiple purposes. In other words, a multi-purpose information system is valuable in terms of data consistency, elimination of redundancies, reduction of information production costs, robustness of statistics. This principle is considered essential for the correct and efficient management of the data acquired by the banks and financial intermediaries.

Governance: the Statistics Committee…

Within the Bank of Italy, the Statistics Committee (hereinafter, "the Committee") examines the information requests presented by the various institutional functions and is responsible for the definition of the basic characteristics of the related new information projects. The Committee has a comprehensive view of the data and acts with a strategic perspective. The needs expressed by supra-national authorities are channeled to the Committee via the relevant Bank's organizational unit and become matter of discussion together with the national needs. In other words, the tasks of the Committee include the broad coordination of statistical needs and the design of solutions with a view to multi-purpose use of data.

securities register. The latter is historically very complete and detailed also thank to the role of National Numbering Agency played by the Bank itself. The data of the entities to whom the information collected by the Bank of Italy refers have been managed, since 1983, in a entities register in which intermediaries, individuals, public and private companies are registered and uniquely identified. This register, in addition to being functional to the numerous surveys carried out by the Bank of Italy, is also one of the sources of the register of institutional units of interest of the ESCB, managed by the European Central Bank and called RIAD (Register of Institutions and Affiliates Data).

… and its role of strategic governance of the common information asset

Precisely for its strategic role, the Committee is chaired by a member of the Governing Board and it comprises the heads of Bank of Italy's Directorates representing the various institutional functions which, for their own analyses and assessments, need statistical information on the Italian banking and financial system. Moreover, the Committee includes also the heads of the Statistical Data Collection and Processing Directorate, which plays the fundamental role of administrator of the common statistical information asset, and of the IT function, which provides the fundamental technological support for the management and use of information. With the support of an operational working group (called the Contact Group), every year the Committee launches an internal survey to gather the new statistical needs[13]. The coordinated action and integrated view of the overall information needs is a guarantee of the final homogeneity and comparability of data and, unlike the silo approach, prevents users from having to carry out costly processes of *ex post* connection between non-harmonised pieces of information. Once all parties agree, the new information needs are channeled along a formal and standardised internal process that culminates with the definition of the issuance of the new Bank of Italy's data requirement. On the one hand, this structured process may appear burdensome and slow, on the other it has proved to be fundamental in order to have an overall vision of the requests and define their priorities. In our view, this process represents an essential element for the correct management of a strategic resource such as statistical information.

The definition of a new information project

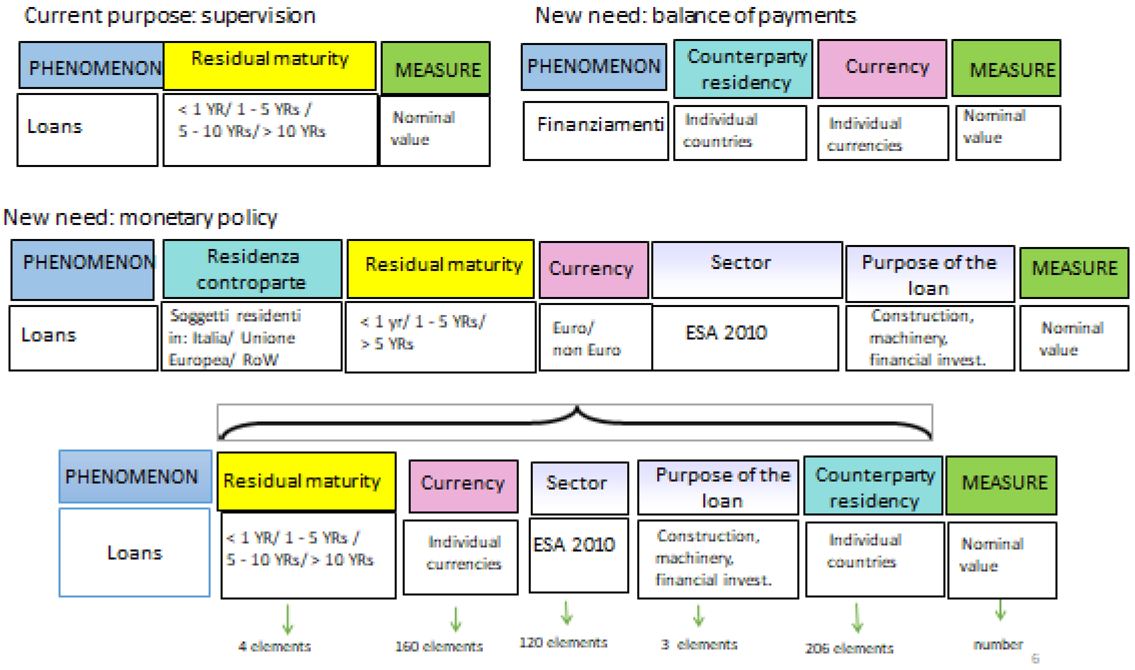
Once the different needs are identified, the preliminary stages for the definition of the new information project begin: it is a multidisciplinary activity that involves various figures. First, the Committee investigates to assess whether the information asset already includes a proxy that is sufficient to meet the new needs. Should that not be the case, it shares the new need with the other users to verify any common interest in the new data and/or gather proposals for possible additions. Third, it carries out a preliminary examination of the potential costs that will affect both the reporting agents and the organizational units of the Bank of Italy responsible for managing the information, by comparing them with the claimed benefits of the new information. To carry out this preliminary and general assessment, the Committee is supported by the PUMA working groups (see par. 4.2). Additionally, the Committee identifies the right level of detail for the new reports balancing, on the one hand, the advantages associated with greater granularity and, on the other, the related additional charges for intermediaries and the Bank of Italy. It also specifies the (internal and external) data circulation criteria and any restrictions to protect their confidentiality. In so doing, it also takes into account the possible impacts of data protection primary legislation. Moreover, the Committee is responsible for the definition of the optimal level of quality controls, with the aim to provide end users with truly appropriate statistics (so-called "fit for purpose") without excessively burdening the reporters. Lastly, it evaluates the need to provide a new IT application or modify existing ones to support new data (see par. 3.3)[14].

Figure 1 presents a purely descriptive example of integration between existing reports (acquired to respond to supervisory purposes) and new information requests that emerge to meet monetary policy and balance of payments production needs.

[13] The survey of statistical projects is annual and refers to a two-year time horizon. The Contact Group updates the project portfolio on a quarterly basis and periodically submits it for review by the Statistics Committee, which assigns priorities and approves project implementation times.

[14] For statistical projects that involve a significant IT commitment, the activities carried out within the Statistics Committee are closely linked with those for the planning of IT projects which falls within the remit of another high-level Committee, the Information and Technology Committee.

Figure 1 – Example of integration between existing and new requirements



The integrated approach is always pursued by evaluating its costs and benefits

We would like to stress that the integrated approach to the information collection is never an absolute goal but it is pursued bearing in mind both costs and benefits. In fact, in some cases the search for elementary reporting capable of satisfying several needs could imply a particular complexity of the reporting model or an excessive granularity of the data; in these circumstances a certain level of redundancy of the reporting requirements in different and separate data collections may be preferable. In such cases, however, redundancy does not mean that the very same data is requested several times. It rather takes the form of separate requests each for a different facet of the same phenomenon. Remaining in the example shown in Figure 1, let us suppose that it is also necessary to collect information on the location of the reporting agent's branch. In such case, it could be more efficient and simpler to define an additional and separate reporting for loans by branch location rather than to extend the classification variables of the same information structure by also including the variable "branch location". In brief, when it becomes too complicated, it is more practical to break down the different requirements (once by "branch location", another instance by "customer location" and, according to the example of Figure 1, by all the other variables) rather than combining all possible variables into a single request for extremely granular data. In practice, these cases are quite rare; however their practical value is sometimes recognised (if not even solicited) by the reporting intermediaries themselves.

The collegial discussion of information needs

In summary, before a new information need becomes a reporting obligation for reporting intermediaries, it is collectively discussed within the Committee in order to ascertain its various profiles and to identify, with the involvement of all the users, the most efficient definition, from the point of view of the cost-benefit ratio of the information to be collected. At this point, the original idea of the new set of information evolves into an actual project that is included in the statistical projects portfolio of the Bank.

3. The importance of a single statistical dictionary for integrated data management

3.1 The dictionary and the data model

The role of the statistical dictionary

The fundamental elements of an information system are a standardised and unitary model of data representation, a catalogue based on a single statistical dictionary, together with a robust IT architecture that ensures, among other things, the logical and physical integrity of the data. These factors enable the information system to: a) support a high level of data processing automation; b) reduce the costs and the time required to design new statistical information; c) allow an easy and flexible matching of the regulatory reports to the evolution of information needs; d) ensure the uniqueness of the information managed, thus avoiding unnecessary redundancies; e) facilitate the consultation of data and the joint use of different information domains, thus avoiding costly *ex post* reconciliations between concepts and definitions; f) allow greater substitutability of data management staff; g) facilitate the learning process of the contents.

The matrix model and metadata

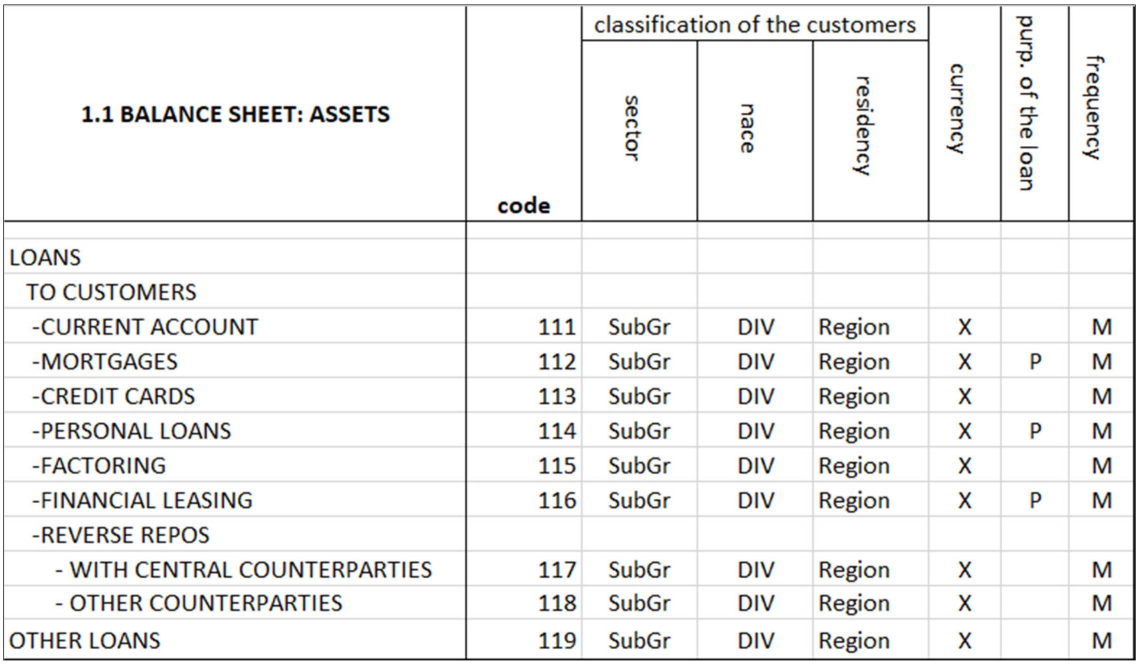
In the Bank of Italy, the representation of credit and financial data is based on the so-called matrix model, which has been adopted since the 1980s and which has developed over time on the basis of the experience gained in the field of statistical information management. This model, or rather its graphic representation, is represented by a double entry table where the rows host the phenomena (e.g., loans and deposits) and the columns show the characteristics that qualify the phenomena (i.e. variables such as the original maturity, purpose of the transactions, residual maturity, geographical location of the customer/branch, the economic sector of the counterparty). Each variable is associated with a list of admissible values, the so-called "domain in use", within a more general domain of values. For example, the variable "residual maturity of the loan" can assume only a sub set of the all foreseen values of the "maturity domain". With the matrix model the observed phenomenon is represented with an n-tuple of values, or with a multidimensional record of the reporting scheme called "multidimensional cube" (see. Del Vecchio, 2007)[15].

In particular, the economic phenomena represented in the rows are described by a numerical code and by a series of variables that detail the phenomenon and which are represented in the columns[16]. Figure 2 shows an example of the matrix model applied to the balance sheet - assets - of a reporting institution.

[15] To deepen the theoretical aspects, methodologies, models and architectures that characterise the statistical information system of the Bank of Italy, see "Statistical Information System" in the Statistics section of the Bank of Italy website.

[16] At the row-column intersection, an "x" is indicated when the phenomenon can assume all the values of the associated statistical domain or a letter when the possible values it can assume are within a domain in use (e.g., in the context of the entire domain of "territorial location" which ranges, according to a hierarchical logic, from Italian municipalities to continents, "R" is the domain in use by the Italian regions, "SG" is the domain in use corresponding to the subgroup of economic activity; in the context of ATECO, "D" is the domain in use corresponding to the level of the Division; "C" is the domain of credit purposes). The periodicity of data submission is, in the example, monthly.

Figure 2 – The matrix model



1.1 BALANCE SHEET: ASSETS	code	classification of the customers			currency	purp. of the loan	frequency
		sector	nace	residency			
LOANS							
TO CUSTOMERS							
-CURRENT ACCOUNT	111	SubGr	DIV	Region	X		M
-MORTGAGES	112	SubGr	DIV	Region	X	P	M
-CREDIT CARDS	113	SubGr	DIV	Region	X		M
-PERSONAL LOANS	114	SubGr	DIV	Region	X	P	M
-FACTORING	115	SubGr	DIV	Region	X		M
-FINANCIAL LEASING	116	SubGr	DIV	Region	X	P	M
-REVERSE REPOS							
- WITH CENTRAL COUNTERPARTIES	117	SubGr	DIV	Region	X		M
- OTHER COUNTERPARTIES	118	SubGr	DIV	Region	X		M
OTHER LOANS	119	SubGr	DIV	Region	X		M

Make complex things simple

In summary, when representing and cataloging the information in a multidimensional model, the final data is the result of the combination of the observed phenomenon and the different variables or characteristics that describe it. In this way, it is possible to represent even particularly complex and detailed pieces of information regardless of their nature, content and specific use.

The advantages of the multidimensional model

It is important to underline that the matrix model follows a general representation irrespective of the specific report layout used in the final analysis. The model, in fact, allows to optimise the collection of data for different end user requests by applying a general multidimensional model which makes it possible to collect elementary information only once and to subsequently recombine it according to different algorithms to produce the various reports (or templates) necessary for end-users. In technical terms this data collection model is also defined "data-driven", as opposed to the one called "template-based" (see par. 3.2). In the data-driven model the data are not collected from intermediaries according to the final layout needed by the user, rather they are acquired "only once" and in a more analytical form so that they can be aggregated in various ways to meet the needs of different users. In addition to what is illustrated in par. 3.2, again as an example, let's consider the case where three users need information on bank loans to carry out their activities, broken down according, each, to a different variable: (1) the customer's location; (2) the sector of economic activity of the borrower; (3) the type of the loan. One possibility could be to ask intermediaries to report loans in three silos, each according to a different detail requested by users. The alternative, which corresponds to the approach followed by the Bank of Italy, is to acquire highly granular information from intermediaries - that is the loans simultaneously by type, geographical location and sector of economic activity of the borrower - and obtain, after appropriate transformations, the three specific reports required. In this way, we avoid duplication in the data production, management and archiving phase and we obtain greater consistency between the information used by the users as they originate from the same elementary data.

In such system, therefore, the user expresses his/her information need. However, the expert in the design of the survey identifies - also by liaising with the same intermediaries and taking into account the data already available - the most efficient method for data collection.

The fundamental importance of metadata

The data-driven model is also the foundation of the statistical dictionary of the Bank of Italy that records, in a unique and unambiguous manner, all the concepts and related definitions. A relational database describes all the data (elementary or calculated) as well as their measures and the structures of the multidimensional cubes. We refer to metadata, as the essential pieces of information to understand the nature, structure, location and content of each individual data stored in the DWH. As an example, the metadata of an information system can be compared with the information used to catalog the books in a library with regard, for example, to the availability, characteristics, main contents, authors, language and the shelves where they are located.

The dictionary contains the transformation rules underlying the DWH statistics

The metadata system backing the dictionary allows transforming the elementary data of the regulatory reports into all relevant derived indicators and statistics. Besides, with the aid of the Dictionary, it is also possible to use the intermediate outputs of the data calculations and even the aggregates used in the automated quality checks that are likewise stored in the DWH. The availability of pre-calculated statistics reduces the computational burden of end users and ensures the consistency of the definitions underlying the data, regardless the use. In any case, elementary data are always available in the DWH so that users, when necessary, can freely process them to meet any specific need.

The registers of entities and securities play a fundamental role in minimizing the burdens on reporting intermediaries

A fundamental role enabling a full application of the integrated approach and of the "collecting once only" principle is played by two peculiar components of the statistical dictionary, i.e. the Entities and the Securities registers (see paragraph 1). In particular, the Entities register contains the descriptive information related to natural and legal persons that are either reporting agents or appear as counterparties of transactions in the regulatory reports. The Securities register, instead, contains the descriptive data of the Italian securities and of foreign securities used in residents' transactions. Each entity and security is identified by its unique code; in the reports submitted to the Bank of Italy, intermediaries are required to report only the code. This simplifies the reporting since there is no need to repeat in every report the characteristics of the entity/security (e.g., the sector of economic activity or residency for the entities and price or issuer for securities). The statistics derived from the regulatory reports are compiled by joining the latter with the descriptive information contained in the registers using the code as a common key.

The advantages of the multi-dimensional model combined with the dictionary

The multidimensional (or matrix) data representation model and the single statistical dictionary are fundamental pillars of the integrated approach to the management of statistical collections from banks and financial intermediaries. They allow, in fact, to use uniform coding methods for all the surveys and for the related derived statistics and to pursue the goal of semantic integration (maximum reuse of the variables in the different surveys) of all the variables. Such an approach makes it possible to reuse the same units of measurement, avoid information redundancies (in case of a new information request, the single dictionary allows the reuse of already existing information domains), ensure intrinsic consistency in all phases of the information production process and, lastly, enables to represent information independently of specific uses. To give a concrete example, a certain aggregation of data used for quality control purposes can also be used to disseminate the same statistics to the end users. In fact, thanks to the dictionary, end users can consult both the values and the related definitions in an integrated way, using the latter as a guide for searching and interpreting the information of interest. In addition, defining information by means of a metadata system is a powerful and flexible solution when dealing with large amounts of ever-changing data. In most cases, in fact, the processing of new data, the modification of those already processed and the introduction of control formulas result in involving just an adaptation of the underlying metadata, without having to intervene on the software (see par. 3.3). This requires a relatively short implementation time and limited resource consume, also because the metadata definition is a task that the data administrators undertake almost without IT support. We can therefore consider the dictionary as being the heart of the management process of the information received from banks and financial intermediaries.

The organisational aspects of the management of the statistical dictionary

With regard to the organisation supporting the statistical process, the management of the single statistical dictionary is based on a semi-centralised and cooperative approach. As a general rule, the operational management of the various information domains of the dictionary is assigned to a single organisational unit. However, the management of some specific information domains (typically the economic phenomenon) is distributed among different organisational units, each taking care of a dedicated segment of the domain. This distribution of work tends to take advantage of local knowledge and skills, only if that helps to better manage the domain.

When a new phenomenon needs to be included in the regulatory reporting, the new element and its code must be added to the statistical dictionary. This operation always requires verifying that no double coding of the same phenomenon is determined and that each new element has its own specific description, which identifies without ambiguity the phenomenon of interest. The single statistical dictionary therefore proves to be a powerful and rigorous tool to prevent spoiling the information system with the introduction of further definitions and variables, when it is not strictly necessary. Provided it does not imply an excessive reporting burden, the typical answer to the request for different facets of the same phenomenon, is to drill down and collect more granular data that can be suitably re-aggregated to produce the different statistics requested by the users. This conceptual rigour is beneficial to the overall efficiency of the use of the data and of its production process, including the quality control of information and the comparability of analyses.

3.2 The differences between the data-driven and the template-based approach

The EBA has adopted a model designed on the final result, the so-called "template-based"

The choice of data modeling made by the EBA as from 2010 to harmonise European banking supervisory reporting[17] follows a different approach. In fact, to express its information requirements, the EBA has opted for a "template-based" data representation (and collection) model on which the Data Point Model (DPM) is developed[18]. Template-based means that the model is designed on the final output (i.e. template, report) to be used by the analyst. At the very beginning the Bank of Italy applied the integrated approach also to harmonise EBA supervisory data to the other national data collections. Therefore, it had initially translated the information content

[17] Among the various harmonised reports that investment firms, banks and, in some cases, also financial companies must produce, the Financial Reporting (so-called FINREP) and Common Reporting (so-called COREP) are particularly relevant. The reporting methods of COREP and FINREP have been defined by the EBA through the Implementing Technical Standards - ITS which represent the set of reporting instructions and uniform schemes at European level.

[18] The Data Point Model is a data dictionary that includes the harmonised information requirements developed by the EBA.

of the EBA's Implementation Technical Standards into the matrix model and national dictionary. The objective was, on the one hand, to limit the reporting burden and, on the other, to benefit from the advantages of the integrated approach. The need to have uniform reporting requirements across countries manifested by international banking groups and also operational risks arising from possible misalignments between the ITS and their transposition in the Italian reporting rules, have eventually led to reconsider the initial choice and opt for the direct adoption of the ITS[19]. This resulted in the inevitable introduction of duplications in the definitions and data structures in the Bank's statistical dictionary for every ITS piece of information which was already coded in Bank of Italy's dictionary.

Instead, the ECB has adopted a "data-driven" approach

It is notable to mention, at this point, that the ECB[20] has essentially opted for a data-driven approach. The different choice adopted by the two European authorities to model the information acquired by banks and financial intermediaries - template-based and data-driven - is currently at the center of a debate. In particular, the discussion is urged by the banking industry[21] which, in detecting the inefficiencies generated by such double framework, hopes for the adoption of a single European methodology for modeling the information requirements and for a single European statistical dictionary for supervisory, resolution and statistical reports[22].

The two models compared in the international debate

In this debate, several arguments are in favour of the functional superiority of the data-driven model. In particular, it should be borne in mind that one of the main goals pursued with the adoption of a common statistical dictionary, based on a single data model, is the absence of duplication in requests for information addressed to intermediaries. This avoids redundancy of data in the information system. If, for example, some pieces of information can be calculated on the basis of other reported data, they should never be part of the data requirement but should be compiled by the recipient authority. Even more so, if the calculations are simple algebraic operations applied to the elementary data of the same data collection. The template-based data representation is, instead, far from this logic since it shifts to the reporting agents the burden of calculating derived information. Whereas, in the data-driven approach, this task is easily performed by the authority responsible for data collection. We observe, in fact, many information redundancies in the template-based model. Some are even quite trivial such as the totals and subtotals of the items contained in the templates. Other redundancies may not be immediately obvious. They are embedded in the validation rules, that is to say, the data quality controls required by the EBA[23].

[19] The experience gained in the early years of the new European supervisory reporting architecture has shown that the choice made on primary reporting raised various critical issues mainly due to two factors:

1. the process of updating the reporting scheme by the EU institutions often left the Bank of Italy and the reporting agents too short (between the publication of the definitive instructions in the European Official Journal and the first effective date of reporting) for respective implementation activities;
2. in other European countries the direct adoption of harmonised schemes and formats is prevalent, so the Italian option of primary reporting based on national rules resulted in additional costs for international banking groups.

[20] See the ECB website for the *SMCube Information model.*

[21] In particular from the European Banking Federation: see *press release of 30 October 2018 and response to the EBA Discussion paper of 16 June 2021.*

[22] The statistical reports referred to in this paper include a large set of data which, at European level, are mainly specified in some regulations or statistical guidelines of the ECB. These include in particular: the reports on the balance sheet data of monetary financial institutions (ECB Regulation (EC) No. 2021/379 of 22 January 2021), those on the interest rates charged by intermediaries (Regulation (EC) 1072/2013 of the ECB of 24 September 2013), the reports on holdings of securities (ECB Regulation (EC) 1011/2012 of 17 October 2012), those of the AnaCredit survey on loans granted by the European banking sector (Regulation (EC) 2016/867 of ECB of 18 May 2016), reports for statistics on payment systems, information on banking intermediaries useful for the compilation of foreign statistics (Guideline ECB / 2011/23 of 9 December 2011 and subsequent amendments), granular reports for money market statistics (ECB Regulation (EC) 1333/2014 of 26 November 2014).

[23] An emblematic example is that of the COREP template C.7 00 in which it is asked to report both the information "Exposure net of value adjustments and provisions" (column 40) and "Original exposure before the application of the conversion "(column 20) and the

In our view, a template-based approach to model data collections was a reasonable solution in the early stages of EBA's activity. That is when the requested data were mostly aggregated, scarcely detailed and collected through Excel sheets (or similar end-user methods). In a historical perspective, we may presume that the reason why in 2004 the Committee of European Banking Supervisors (i.e. the European authority from which the EBA was originated in January 2011), opted for a template-based data collection format, is because they did not feel the need to set up an organizational unit specifically dedicated to data administration. In fact, the template-based data collection has the advantage of requiring a minimal data administration since the information is already displayed according to the users need.

However, the rise of increasingly granular surveys has made the template-based approach inefficient. Eventually it has become evident that end users' visualisation needs cannot represent a constraint on the (technical) way chosen for data collection. Otherwise, duplications in the reporting obligations are inevitable. Of course, users should always put forward their information needs but data managers must be able to decide on the most efficient method of collecting elementary data, as long as they ultimately ensure that the various necessary layouts and data transformations are applied. Since the statistical dictionary also contains the metadata rules to transform elementary data into more complex pieces of information (indicators, relevant aggregates, other statistics), with a data-driven approach users are always able to obtain the necessary level of aggregation and transformation, without thereby spoiling the integrity and efficiency of the data collection model.

3.3 The technology supporting the data collection

In the Bank of Italy, IT applications to support data management follow the integrated approach …

The Bank of Italy has developed a dedicated IT platform to support the management of the data collected from banks and financial intermediaries and its subsequent use. Even the IT tool is seen as a fundamental element of the integrated approach so much that it has been designed consistently with the general logic of the production of statistics. In fact, to ensure a good quality level of the extensive information asset, of the Bank of Italy, it is of fundamental importance that the data acquisition and processing phases 1) automatically incorporate the decision-making processes of the data managers ; 2) allow a swift and safe data exchange with the reporters; 3) allow the continuous versioning of each data point ; 4) support the Bank of Italy's accountability by keeping track of the results of all calculations and checks. Furthermore, it is of the utmost importance to count on a rapid change management to ensure both a rapid adaptation of the DWH to host new information and that its evolution is in line with international data management standards.

… assisted by the adoption of metadata-driven software

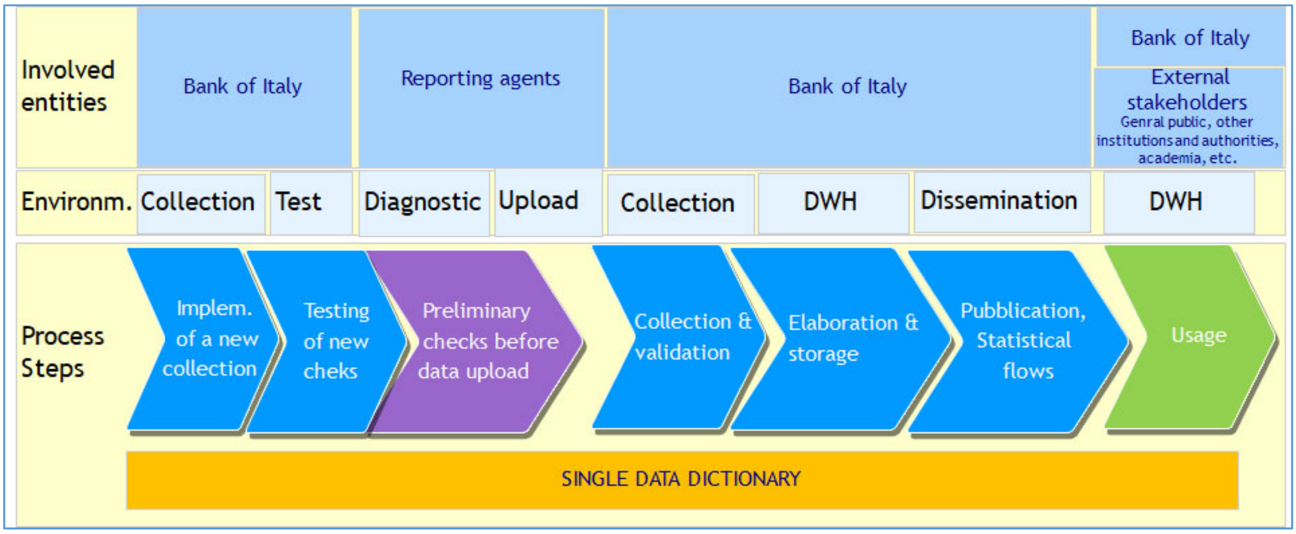
In particular, the Bank of Italy has opted for the creation of an integrated technological platform, which organically supports the phases of data collection, processing and dissemination. This solution is metadata-driven, which means that the calculation services are defined "only once" and are orchestrated through a metadata system – based on common, homogeneous and shareable rules - that allows to customise each step of the process and to keep track of every calculation. The parameters guiding the underlying software are, in fact, metadata themselves described

"Value adjustments and provisions associated with the original exposure "(column 30). It is quite clear that the information in column 40 can be calculated through a simple mathematical difference between columns 20 and 30. To testify this mathematical relationship there is also a validation rule that the reporting agents must comply with.

in the single dictionary. Data managers administer the metadata independently from the IT. They allow the software to automatically carry out, all the desired processing (e.g. data control, archiving in the DWH of the validated flows received from intermediaries, transformation of elementary data in intermediate and final statistics, dissemination to internal and external users). Through the metadata system, it is possible to trace the entire path followed by the data and reconstruct the value taken at each stage of the calculation process (versioning). The metadata therefore constitute the engine of the production chain of statistical information and, therefore, the very heart of the information system.

This integrated and metadata-driven IT solution makes it possible to support end-to-end the life cycle of the data as a single flow of consecutive actions (see Figure 3). With the aid of a control point system, a dedicated organisational unit monitors the successful progress of the overall process through each stage. The process is divided into two phases: a preparatory phase (i.e. the testing of new control algorithms, the implementation of the characteristics of new information segments, the definition of new controls and the setting of the calculation of indicators and aggregate statistics) and a running phase (the collection and validation of dataset updates, the periodic processing of aggregates and indicators and the internal and external dissemination). In turn, all the environments hosting a segment of the process (i.e. test, collection, validation, DWH and dissemination) are defined on the basis of a generalised and unique information model. This integrated model can deal with different categories of structured data (qualitative, quantitative and, among the latter, time series and cross-section data), variously complex rules of aggregation and transformation, logical interdependencies between information (so-called hierarchies) and, last but not least, is able to set an automated dialogue with the reporting agents on the outcome of the quality controls.

Figure 3 – The phases of the data management process and the related environments of the IT platform



The IT system is applied flexibly to manage occasional information needs

The IT system, however, is flexible enough to also allow for the management of extremely urgent and/or occasional information data needs which may require a fast setting, thus offering the possibility of skipping some of the steps foreseen for the regular process described above (fast track). The IT procedure, in fact, can accommodate the quick implementation of occasional data collections without sacrificing the provision of a full data integrity and security backing the standard surveys. Should such occasional information requirement become a regular data request and should that be of common interest across multiple functions in the Bank of Italy, the data management will then shift to the full process described above.

4. The different dimensions of data quality

The whole process is aimed at ensuring data quality

After the evaluation step of the new information needs and the design of the data necessary to satisfy them, the process of collecting the new information comes to life. The next steps are the issue of the reporting regulations, the dialogue with the reporting agents on how to best comply with the requirements, the actual data production and submission of the reporters and, lastly, the collection and control of data by the Bank of Italy. The activities carried out in all these phases are arranged so as to guarantee the highest quality of the reports, the latter being an essential prerequisite to ensure sound analyses and decisions of the data recipient.

In multiple occasions and in line with the general principles established for statistical reporting by the ECB, the Bank of Italy has highlighted that information takes on its full value only when the data are fit for purpose. This means that meeting high data quality standards is a fundamental goal to maintain, and possibly increase, public trust in the official statistics produced by the Bank and on which grounds policy decisions are taken.

The different profiles of data quality

For the purpose of this paper, the concept of data quality is expressed in terms of the "accuracy" with which the data represent the observed phenomenon, the "consistency" shown between cross-linked information segments, over-time and cross-section "comparability", the “relevance” of how data cover the users’ needs, the “timeliness” for the users to obtain the data. Another important dimension is “cost-effectiveness”, that is the comparison between the costs of statistical production and the expected benefits for the users[24]; this is pursued in the intense dialogue with the reporters.

We discuss about “relevance” and “cost-effectiveness”, respectively, in chapter 2 and par. 4.2. With reference to the other dimensions of data quality - accuracy, comparability, consistency and timeliness - the tools on which Bank of Italy leverages relate to different profiles: (1) clarity, completeness and timeliness of the reporting instructions; (2) reliability of the processes for compiling the information flows submitted to the Bank of Italy by the reporting intermediaries; (3) specialisation in the data collection and validation process. In the remainder of the paragraph these issues are examined in depth and it is highlighted how all the actions lead towards the construction of an overall quality system that is not limited to the *ex post* control phase of the report received - although

[24] For the dimensions of the quality of the statistics, principles 11 to 15 of the “Code of European Statistics” adopted by the Committee of the European Statistical System in 2017 are used as a reference (see Eurostat (2017)). In the context of the ESCB, similar principles are contained within the “*Public Commitment on European Statistics by the SEBC*” (https://www.ecb.europa.eu/stats/ecb_statistics/governance_and_quality_framework/html/escb_public_commitment_on_european_statistics.en.html). At the international level, the reference system on the quality of statistics is the one adopted by the United Nations and last revised in 2014, whose principles are also inspired by the quality standards defined by the International Monetary Fund provided for by “*Data Quality Assessment Framework*” (see IMF - 2003).

obviously necessary - but concerns all the phases of the collection process and involves reporting agents as well as authorities.

A further dimension of data quality, of which we provide only a few hints in this paper, concerns "integrity". This concept can be viewed either as (1) the guarantee that the information is not altered as a result of errors, voluntary actions or malfunctions of the technological systems or (2) the guarantee that the statistical production is based on high professional, transparent and ethical standards[25]. With regard to the first meaning, we are of the opinion that a metadata-driven software acts as a facilitator in the application of the integrated approach because, among other things, it keeps track of the various calculation steps and allows for a sound data versioning (see par. 3.3). When it comes to the second aspect of integrity, we all know that in the central banking world the term has a much broader scope. It evokes ethical values and principles that should characterise every activity. Being information a sensitive asset, such general scope is indeed strengthened by the adoption of corporate principles and policies established by the top management and uniformly applied by the various organizational units involved in the data production process.

4.1 The importance of clear and timely reporting instructions

Effective interaction with the reporting entities is essential

In the preparatory work to provide the Bank of Italy with quality data, a leading role is played by the producer of the basic information, i.e. by the reporting intermediary that provides the elementary data on which indicators and statistical aggregates are compiled. It is therefore important that reporting institutions are put in the best conditions to fulfill their role.

It is essential to announce in advance the projects for new surveys ...

First, it is essential that the authority plans and informs well in advance about the new statistical projects that will have an impact on the reporting entities. In the past, the Bank of Italy used to issue an annual specific communication plan to inform reporters about the innovations in statistical and supervisory reporting planned for the following two years. The Bank had to abandon this practice when it joined the ESCB, the ESFS, the SSM and the SRM, that is, when the evolution of national information requests became to be closely intertwined with the European reporting legislation and therefore they had to follow the deadlines there established. However, also in the new scenario, the fruitful dialogue established in the "PUMA cooperation" (see the next paragraph) between the Bank of Italy, the banks and financial industry, and their respective trade associations, allows in any case, to ensure an important and timely circulation of information on the emerging European reporting initiatives.

… as well as to issue clear and timely reporting rules

In order to enable the reporting agents to carry out the necessary in-depth analysis of the reporting rules and then correctly implement them in their systems, it is also crucial that the legislation regulating the details of the data reporting is issued in clear and timely fashion. The national reporting instructions directed to banks and non-bank financial intermediaries have been combined over time into specific regulatory texts (so called Circulars) and are published on the Bank of Italy's website[26]. They are issued well in

[25] International Monetary Fund - Data Quality Assessment Framework: – *Integrity - statistical policies and practices are guided by professional principles; statistical policies and practices are transparent; and policies and practices are guided by ethical standards.*

[26] Circulars are disseminated in the section Statistics/Legal framework of the Bank of Italy website. As already highlighted in paragraph 3.2, with reference to the supervisory information required by the EBA and to the Implementing Technical Standards (ITS), the Bank of

advance to reporters (usually from six months to one year before they enter into force, although in some cases the timing is negatively affected by the time needed to finalise the European regulations). Circulars bear all the details of the reporting obligations applied to banks and financial intermediaries and, where necessary, they simply make reference to the European legislation without any duplication or overlap. When applicable, the Circulars also include clarifications, agreed at European level, in response to specific questions presented by the intermediaries themselves. Usually the Circulars are complemented with some practical examples and their annex includes reporting schemes and coding systems. The clarity and level of detail of the Circulars also benefit from the input received within the PUMA cooperation initiative. In this way, the need to update the legislation is considerably reduced, although the texts are enriched and refined over time as a result of further inputs received from reporting entities.

The ultimate goal of the Bank of Italy is to regulate reporting obligations by means of an organic, clear and precise secondary legislation framework, in order to minimise possible misunderstandings of rules and definitions since this would ultimately penalise the quality of the data collected.

A correct communication to data producers that contributes to the quality of information is also comprising:

Another element is the transparency on controls and uses of the data

- transparency on procedures to verify the quality of the data. The Bank of Italy (1) shares with reporters the main types of quality checks applied to the reported data; (2) provides them with specific diagnostic software to identify possible errors prior to the official transmission of the data; (3) discloses the formula behind the checks when errors are detected in the reported data;
- transparency regarding the use made of information. The Bank of Italy's website describes the main purposes pursued with the data reported by intermediaries.

4.2 The PUMA cooperation initiative

Collaboration with regulated entities from the earliest stages of production of the flows ...

To have good quality data it is important to monitor the production of information from the earliest stages. The relevance, sensitivity and complexity of the context in which the Bank of Italy operates, such as that of regulatory reporting, demands the establishment of a robust data production process right at the regulated entities. This is, recognised by Bank of Italy's Circular n. 285 on the supervisory provisions applicable to banks and Italian banking groups. In particular, the chapter concerning the information system specifies that *"when using the company data warehouse for analysis and reporting purposes, the procedures for data extraction, transformation, control and loading in centralised archives - as well as the data exploitation functions – should be documented in detail, in order to allow verification of data quality. The reporting system should allow to produce timely and high quality information for the supervisory authority and the market"*.

Knowing, guiding and contributing to set the processes that the data producers use to compile regulatory reporting allows the regulator to achieve multiple goals. In particular it helps to: (a) avoid non-homogeneous interpretations of the reporting instructions among intermediaries; (b) prevent situations where the reporter's data production is affected by problems concerning the data extraction and calculation; and (c) ensure consistency between different information flows, each aiming at satisfying different

Italy, after an initial period in which the EBA ITS standards had been converted into the proprietary matrix model, preferred avoiding a misalignment and adopted directly the European Regulations and the EBA technical instructions.

purposes, which are nonetheless based on the same company raw data (an example is represented by monetary statistics with respect to Central Credit Register information). Overall, these actions help to maximize the compliance with the data quality dimensions described at the beginning of this chapter.

… takes place in Italy with the "PUMA cooperation" initiative

In Italy, to achieve these purposes, a fundamental role is played by the structured cooperation between intermediaries called "PUMA" (*Procedura Unificata Matrici Aziendali*; i.e. unified procedure for the company matrix reporting), promoted and coordinated by the Bank of Italy[27] since the late 1980s. The main tangible output of such cooperation is the PUMA documentation that describes, in a formalized language based on metadata, the rules for the production of regulatory reports by banks and financial intermediaries and guides the extraction of the raw information stored in the reporting agent's archives[28]. Without this coordinated effort, each reporting agent would individually decide which exact raw data must be extracted from the company archives (the so-called input data) and which specific algorithm should be then applied to produce the reporting obligations. This hinders the homogeneity of the application of the regulations across reporters. In this regard, it is important to note that, despite it being produced by the voluntary members of the PUMA groups, the documentation is published on the PUMA cooperation website[29] and therefore is freely available to all reporters.

The purpose of the PUMA cooperation

At its onset the PUMA cooperation exclusively involved bank. It was later extended to financial enterprises engaged in leasing, factoring, consumer credit and guarantee issuance. It was launched by the Bank of Italy with a dual goal: 1) increase the quality of the reports by favouring the homogeneous application of regulations and reporting instructions; 2) improve the awareness of intermediaries about the wealth of information at their disposal for business management and the related importance of ensuring an efficient organization of their databases. Tommaso Padoa Schioppa, at the time Deputy Governor of the Bank of Italy, had significantly remarked this last aspect since the early 1990s[30]. He, in particular, stressed that the PUMA documentation should be viewed as a tool capable of offering the intermediary the possibility of turning the effort to produce the regulatory reports to its own advantage. In this respect, the PUMA procedure is more than just the reports it generates, because the large database that it makes available to the intermediary constitutes a very rich information asset ready to be also exploited for internal business management purposes. These considerations are particularly meaningful still to this date.

The longevity of this initiative is therefore the result of the concurring interests of the Bank of Italy and of the banks and financial intermediaries (the quality of the data and the exploitation of the rich information assets for corporate governance). By establishing a direct connection between the inner archives of financial intermediaries and the reports to the Bank of Italy and guiding the processes of extrapolation of data from company information subsystems, PUMA still represents the main information system that ensures that reported data are consistent with the reporting rules (including those then directed to the European authorities).

[27] The Bank of Italy chairs the Steering Committee of "PUMA cooperation" and coordinates the related functional banking and financial groups in which 18 banks, 1 intermediary registered in the list referred to in Article 106 of the Italian Banking Law, 5 trade associations in the credit and financial sector, the Cassa Depositi e Prestiti and the Bancoposta Division of the Italian Post Office currently participate, on a voluntary basis. In particular the organisational unit in charge of the coordination is the Statistical Data Collection and Processing Directorate, the same dealing with the banking and financial data collection, compilation and dissemination.

[28] Currently, the PUMA documentation covers a wide range of information domains, which include, inter alia, monetary and financial statistics for the ECB (e.g. AnaCredit and Securities Holding Statistics), the Central Credit Register, the annual financial statements of banks, supervisory and resolution reports.

[29] www.cooperazionepuma.org.

[30] See Padoa Schioppa (1993).

PUMA is not software and its use is voluntary

Sometimes, and even among insiders, there are some misunderstandings about the role of PUMA that we would like to clarify here. First, the documentation is not software and, in fact, intermediaries must implement their own IT solutions to apply the metadata of the PUMA documentation in order to extract the data from their archives. Second, its use is strictly on a voluntary basis and it has no certification value; in other words, intermediaries remain fully responsible for the accuracy of the data transmitted to the Bank of Italy. Lastly, the input data used to elaborate the regulatory reports are not shared with the Bank of Italy.

The advantages of the PUMA approach

PUMA cooperation has important advantages. First, a uniform application of the reporting rules facilitates the achievement of greater consistency in reporting and a lower burden – not only operative but also economic - for reporting agents. The latter is a very sensitive aspect to reporters, especially in consideration of the fact that reporting regulations are increasingly heterogeneous and complex and constantly evolving. Second, the interaction that takes place in the PUMA working groups, specifically given its informal level, guarantees a continuous exchange of information between authority and reporters, which makes it possible to detect possible reporting problems before they actually occur. With regard to the last point, it is important to draw attention on the advisory function to the regulator of PUMA cooperation, albeit informal. In particular, the discussion on the new regulatory requirements allows assessing their impact, also in terms of costs; this dialogue is fundamental to identify the most efficient way to fine tune the information requested and better shape its characteristics[31].

4.3 The collection process and the data quality controls

This section completes the analysis of the production process of new statistical information by focusing on the phase of data collection from reporting intermediaries and their subsequent validation.

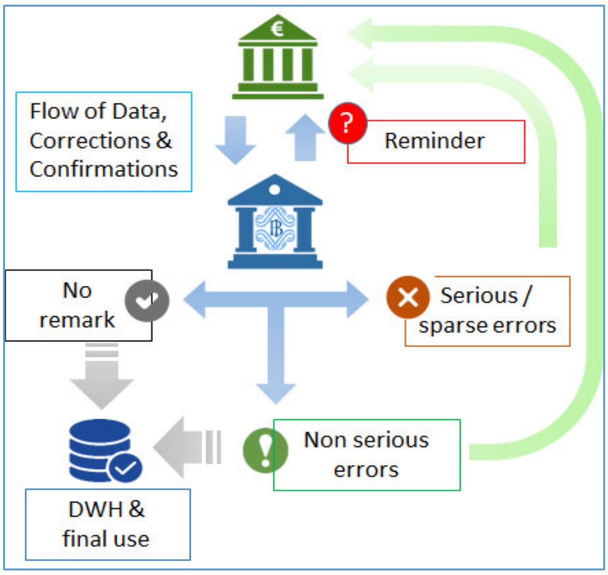
The data arrives in the Bank of Italy

On the basis of a production process founded on PUMA documentation, reporters periodically send their data flows to the Bank of Italy in order to comply with the various reporting obligations (national requirements, Central Credit Register, statistical balance sheets, interest rates, AnaCredit, etc.). The transmission requires to upload the data files onto a dedicated IT platform (see par 3.3) where reporting agents can also conduct many quality checks prior to the official submission. The availability of the diagnostic environment, which includes most of the controls foreseen for each dataset[32], allows the reporting agents to identify in advance any anomalies that are otherwise detected by the Bank of Italy once the file is received. In this way, the anomalies can be, as much as possible, solved before the official sending. As shown in Figure 4, the Bank of Italy uses its IT collection platform to send to the reporting institutions reminders, in case of delay in sending the data, and remarks with the outcome of the checks, if any alleged anomalies are found.

[31] A further input for the definition of clear reporting regulations is derived from the results of the public consultations that the Bank of Italy is required to carry out when new or amended information requirements - regardless of whether they are of national or European origin - entail new charges for the reporting agents. In the context of the ESCB, this involvement also takes place in the context of the "merits and costs procedure". The latter is a procedure expressly provided for by Regulation (EC) 2533/98, and is preparatory to the introduction or substantial modification of European statistical regulations. The procedure is aimed at ensuring that the introduction of new statistical requirements is sufficiently justified by policy needs and is at the same time oriented towards identifying balanced solutions from the point of view of costs and efficiency, useful for minimizing the reporting burden.

[32] Controls that require comparison with other sources are excluded (cross-checks between datasets sent by different reporting entities, cross-checks with internal sources).

Figure 4 – The data flow between reporting agents and the Bank of Italy



Quality controls

The Bank of Italy has put in place a series of measures aimed at ensuring high quality standards. Among the most challenging goals and with reference to the different dimensions of quality mentioned above, a notable mention is the need to reach a correct balance between the timeliness in the availability of information and the achievement of sufficient quality levels. It is almost physiological, in fact, that when the information is transmitted by the reporting agents to the Bank of Italy, it may require a stabilization period during which all the quality checks are carried out and the intermediaries are contacted for the necessary clarifications/corrections. The goal of this sensitive phase is to make sure that the quality of the information released to the users is adequate for their analysis and research. No matter how well designed the information was and how robust and well tested the production processes are, each data flow can still be affected by imperfections. There can be many reasons for such errors and the methodologies used to identify potential outliers are, in fact, constantly evolving. In the following, we would like to present the main features of the approach followed by the Bank of Italy to check the quality of the data received from banks and other financial intermediaries.

The first aspect is the wide articulation of the set of controls to check the quality of the data uploaded to the collection environment. The various checks are automatically activated upon arrival of the data file and they can be classified into four macro categories:

- *format checks*: they verify the correctness of the file with regard to the expected format standard. In case of error, the file is rejected as it cannot be processed. The metadata describing the format checks are automatically generated by the IT procedure when implementing the characteristics of the new data collection. For example, the expected format is SDMX[33] but the actual data flow is an Excel file; the machine does not recognise the file and rejects it.

[33] The SDMX (Statistical Data and Metadata eXchange) is a standard for the exchange statistical data and metadata (https://sdmx.org)

- *formal checks*: they verify the compliance of the coding of each record with the dictionary elements described in the reporting instructions. For example, if loans are expected to be detailed by "subgroup of economic activity" of the counterparty, any reported code that is not part of this domain of values (they either do not exist or are included in other subdomains of the "economic activity" overall domain) will result as incorrect. In order not to waste information, only the record specifically affected by the error is rejected and, where possible, a predefined algorithm replaces the wrong element with a conventional value. This allows to use the quantitative information while waiting for the reporter to correct the data. Also in this case the checks are automatically generated when the information manager implements the metadata that describe the characteristics of the dataset to be acquired;

- *deterministic checks*: they verify that the data file contains all expected data points/records and that the reported values comply with the expected logical relationships between phenomena and related variables. They operate by comparing different values of correlated phenomena reported in the same dataset or values provided by different intermediaries with similar characteristics (peer-to-peer comparison). For example, the total balance sheet assets and liabilities of a reporting agent should match within a given tolerance. These controls are implemented in the metadata system by the information manager in the implementation phase*;*

- *statistical-probabilistic checks*: they verify the reliability of the trend shown by the time series. They are particularly useful when no obvious *ex ante* logical relationship linking phenomena and their variables exists in the data that can be verified with the deterministic checks. Statistical-probabilistic checks are defined by the information managers and are as much as possible customised on the type of reporting agents. They usually imply the use of a tolerance threshold that is subject to periodical assessment. This ongoing maintenance is necessary to ensure effectiveness and efficiency of the control system over time; however, it is particularly burdensome in terms of time and human resources. For this reason, we believe in the strategic role played by innovative statistical methodologies, which allow, among other things, to deal with very complex algorithms and to dynamically adapt the formulae with the very same data forwarded by reporters (see the box "The application of machine learning methods to data quality controls").

Deterministic and statistical-probabilistic checks are associated with different levels of error severity to which as many actions correspond. In fact, the presence of "very serious" anomalies prevents the dataset to be released in the DWH as it remains in the collection environment pending the necessary corrections[34]. Especially if the data file is received well in advance of the deadline, the information manager may decide not to release the data, pending corrections, when the file is affected by minor but rather widespread anomalies. Should the anomalies be only few and "not serious", the data files are released to the DWH immediately after the control phase and, at the same time, the possible remarks are transmitted to the reporting agents to request a correction/confirmation of the data.

[34] Where necessary, the information manager can always manually unlock the release to the DWH of a blocked dataset.

The Bank of Italy closely monitors the observance of the reporting deadlines and the timeliness to solve “very serious” remarks. As already mentioned, in the absence of "very serious" anomalies, the data are immediately released to the DWH; internal users can always visualise the rationale of any pending non-serious remark and evaluate whether the data is nonetheless suitable for the specific intended use.

Specialised organizational units are dedicated to data quality checks

Another relevant aspect to be taken into account when analying the management of data quality within the integrated approach is the organisation. Specialised organisational units, separate from the users and located within the same functional structure (i.e. the Statistical Data Collection and Processing Directorate) are in charge of the data quality process. This approach has the advantage to allow for economies of scope with regard to the statistical skills needed to control the data. Such organisation does not prevent the users to be part of the control chain since their deep knowledge of the phenomena can provide an important input for the *ex ante* definition of new and more specific checking algorithms. Cooperation between data managers and users remains a key element to create high quality databases.

Quality is a goal that is achieved with successive iterations

In our experience the quality level of the data submitted by reporting agents requires a period of time to settle. After the first submission it gradually increases, along with the corrections/confirmation of the anomalies sent by the reporters as a result of an intense interaction with the Bank aiming at correcting the data or receiving explanations on unusual trends, comparing the trends to the more general economic and financial pattern, comparing different cross-linked information flows. As already said, it would be unrealistic to expect the data to be completely free of anomalies from the first submission; in fact, subsequent adjustments are physiological and, within certain limits, do not constitute a symptom of incorrect reporting behavior. The post submission data review process is therefore a fundamental step to achieve optimal data quality.

Quality checks take into account the uses of the data

Lastly, a proper definition of the quality checks also requires to take into account both the actual use of the data, in order to shape the checks on what it is most necessary, and the costs associated with extremely detailed checks. In fact, data managers must always seek the right balance between the desired quality level and the burden of a too analytical nature of the controls. The latter, if excessive, leads to a very high number of potential anomalies that reporting agents have to analyse which, in turn, inevitably delays the moment with no pending remarks since everything is either correct or confirmed. Besides, if not absolutely necessary, authorities should refrain from asking to check the correctness of very old data, because such verifications usually imply very costly and cumbersome procedures for intermediaries (for example, after merges and acquisitions some details may not be easy to trace anymore).

Controls defined at European level on harmonised statistics

So far we have largely described the quality control approach that the Bank of Italy applies to the data falling under its direct governance. However, the progressive harmonization of statistical collection in Europe entails likewise the progressive convergence of national control methodologies. An example are the validation rules applied to the harmonised supervisory reporting defined by the EBA and the SSM supplementary checks defined by the ECB together with the national competent authorities. A minimum set of common validation checks has also been introduced by the ECB to assess the completeness and consistency of AnaCredit data. These checks are defined by the Statistics Committee of the ESCB and are publicly available. Such list, however, does not prevent NCBs from applying additional in-house controls, for example cross-checks with other databases available at national level (e.g. AnaCredit and the Central Credit Register).

The application of machine learning methods to data quality controls

The volume of the collected information is growing along with the interest of the users towards increasingly granular data and wider and more detailed business areas of banks and financial intermediaries. Technology nowadays offers greater possibilities to elaborate big volumes of data and to adopt new and more powerful statistical methodologies. These factors demand a change of pace in the approach adopted in data quality control.

Back in the 1980s the automation of the various stages of the statistical production process (including data quality) represented an important step ahead that allowed significant efficiency gains, because the machine was finally in charge of all predefined repetitive manual operations, including data verification and comparison.

Today it is even possible - and efficient - to delegate to the machine the design of the controls and part of the periodic overall assessment of the control system. Similarly, the machine can replace the data manager in some of his/her decisions. These are the areas where in-depth studies are underway in the Bank of Italy with the aim to adopt innovative statistical methodologies that belong to the branch of artificial intelligence known as *machine learning*.

It is important to consider that the potential outcome of *machine learning* applications to quality checks goes well beyond the possibility of using very complex algorithms. In fact, a particularly innovative aspect is that the machine is able to follow the evolution of the reported phenomena and consequently dynamically adjust the existing checking rules. This is done through a learning process that uses the outcome of checking formulae applied to the data over time. In this way, it is possible to increase the precision of the checks because they are purposely calibrated to the specific characteristics of the collected data.

There are several advantages of the application of *machine learning* techniques to data quality processes. First, they enhance the overall efficiency of the process because they significantly limit the area of manual intervention, as the machine takes charge of an increasingly large share of the decision-making process. Second, they allow for a greater accuracy of the checks by reducing the so-called "false positives" and, as a consequence, they also contribute to limit the checking burden at the reporting agents end. Third, they improve the quality of the data by quickly identifying anomalies which would otherwise remain unnoticed (so-called "false negatives"). That is particularly relevant when there are no *a priori* deterministic relationships in the data that can be put under control. Lastly, in some cases, the outcome of the advanced checks or calculations may also be used to enrich the information asset already available by providing further statistics and indicators. For example, new quality indicators may be drawn on the basis of the combined analysis of structured and unstructured information (see Chapter 6).

The application field of these methodologies is potentially very broad and destined to expand. In fact, the growth rate of these statistical methodologies and that of technology is impressive. Additionally there are more and more new unconventional sources that can be used with the structured data produced by intermediaries to create new informative value (*databases* available from commercial suppliers or other entities and institutions).

The Bank of Italy[1] has developed several projects of *machine learning* techniques application to improve the processing of data acquired by banks and financial intermediaries. They range from the enhancement of the quality control system on reports, to the automation of the analysis of the confirmations produced by the reporting agents to anomaly remarks[2]. Specific natural language processing algorithms have been successfully applied to automatically extract useful information from the Italian Stock Exchange Notices and use it to enrich the Securities register and to appropriately classify the nationality and type of companies archived in the Entities register[3]. Further works are underway with the aim to further improving the efficiency and effectiveness of the quality checks and to maximise the automation of the different phases of the statistical production process.

In this new context, we observe a shift in the skills needed to perform the quality control task. In the past, we needed operators able to run multiple checks based on conventional logic. Data managers are now progressively more dedicated to design new innovative and powerful statistical methodologies that are able to scan in-depth the quality of large volumes of data with little operative effort.

(1) The first step was taken in the 2017-2019 Strategic Plan, point 3.1.b "develop and test innovative statistical techniques (such as machine learning and big data), also by activating collaborations to evaluate opportunities and risks".

(2) E.g. Zambuto F., Buzzi M. R. et al. (2020); Cusano F., Marinelli G., Piermattei S. (2021); Zambuto F., Arcuti S., Sabatini R. et al. (2021).

(3) E.g. Giudice O., Massaro P., Vannini I. (2020); Bernardini M., Massaro P. et al (2021).

5. Data dissemination to internal users

The statistical DWH…

For the Bank of Italy to effectively undertake its multiple and detailed tasks, it is necessary to provide its analysts and researchers with a large information asset and to enable them to extract information with easy-to-use and well documented procedures. A deep knowledge of the information content and of the most appropriate IT tools to facilitate data elaborations become of pivotal importance. Response time also plays a significant role when dealing with millions of data. Besides, the joint use of different information domains must also be facilitated.

… managed according to a centralised organisational model and…

To meet these needs the Bank of Italy has arranged its information asset within an integrated and centralised information system whose central point is the unique DWH. This is, in turn, managed according to a centralised organization model. In other words, as the Bank of Italy stayed away from a "silos" approach to shape the collection phase, likewise this logic was not adopted for the dissemination phase. As mentioned earlier, the application of the integrated approach is not blind and absolute. In fact, the organisation of the information encompasses the possibility to arrange private data marts[35] when the relevance of the information is only limited to a local use. In this case priority is given to the extreme flexibility and quick adjustments in the data management.

[35] The data mart is a local database while a data warehouse is typically an information system available to the entire company.

… developed with a view to integration

The first cornerstones of the system were placed in the late 80's when the specialised literature on database organisation was still quite limited and when, for their calculations, users were still heavily relying on IT support. One element of the integrated approach is that users, and even data managers, in their day-to-day ordinary job must be independent from the IT function. In this regard, a metadata driven software and applications for an easy navigation in the DWH are key. The data collected from banks and other financial institutions and the results of the data quality checks are stored in the DWH. It is also possible to keep track of the different values taken by the same piece of information after corrections (versioning). Most of the data is stored permanently in the DWH, in fact extended retention periods are necessary due to proven institutional needs. Anyway, in compliance with *General Data Protection Regulation* (GDPR)[36] the retention period of the different data segments and the access rules are continuously reviewed.

The integration of external sources

Some segments of the DWH host data coming from external sources like commercial data providers (e.g. financial markets data) or other national and international authorities (e.g. information flows from the ECB). Besides, also the registers are part of the DWH.

The various information segments of the DWH follow the same data model (the matrix model) and are supported by the same single dictionary; the latter also acts as a catalogue to find information. Such integration enables to cross-use the data, be that reported by banks and financial institutions or coming from other sources. It is also used to build meaningful statistics from reports on individual entities by using the reference data (i.e. descriptive attributes) available in the Entities and Securities Registers.

The key characteristics of the DWH architecture are: (1) data integration, as the result of the application of one standardised representation model that is also independent from the data transmission format (e.g. SDMX and XBRL[37]); (2) possibility to apply simultaneous queries on different information segments; (3) elementary data and pre-calculated statistics are both available; (4) redundancies are eliminated; (5) the system quickly adapts to the evolution of the data; (6) long time series are available (in compliance with the privacy legislation when it comes to personal data).

Lastly, an integral component of the DWH is the guided navigation tool, which enables the user to make complex inquiries without the need to have an advanced knowledge on the data architecture.

The data access policy is based on the need-to-know principle

As for the internal access policy, the Bank of Italy has fostered a broad sharing of banking and financial data among the different users, regardless the institutional function and the location (head office or branches). Since banking and financial intermediaries data are designed to be multi-purpose and, in fact, are governed in a cooperative way by the Committee, Bank of Italy's staff which requires for its tasks to access such data, is given full access to the DWH (need-to-know principle). According to the described framework, there are no segmentations by topic: exceptions apart, the various information areas are freely available to all institutional functions.

As mentioned in chapter 2, the composition of the Committee, encompassing all institutional functions that use data collected from banks and other financial intermediaries, allows for a shared governance and an efficient design of the reporting

[36] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016.

[37] XBRL (eXtensible Business Reporting Language) is a standard for exchanging balance sheet and financial information.

obligations and, more generally, of the entire information asset. It also favours the multi-purpose use of statistical and supervisory information within the Bank of Italy.

We believe that an integrated approach to statistical data management can dramatically improve the value of the strategic corporate asset represented by its information wealth. Benefits can be obtained in terms of (1) quality, non-redundancy and great level of detail of the information available to users; (2) efficiency in the data management; (3) lower reporting costs for intermediaries (than with the silos approach).

6. The challenges ahead in a complex and dynamic scenario

The objectives underlying future challenges

In the coming years the Bank of Italy will have to face different challenges pertaining to the articulation of the data collection process from banks and financial intermediaries and to the efficient management of the acquired data in its information system. Some issues are matter of discussion in the various international tables where the Bank cooperates with other European authorities and institutions, others have a purely domestic dimension.

The expected challenges are intertwined and share the same general goals. First, a continuous improvement of the statistical services to better meet the users' needs, in terms of information content and data management methods. Second, the reduction of reporting burden also by virtue of a greater dialogue and coordination among authorities. Third, the evolution of the information production process towards an ever greater robustness, efficiency and economy, by also fully exploiting increasingly complex, diversified and granular data with innovative statistical methodologies (machine learning). Lastly, the construction of a robust data governance, both within individual authorities and at European system level.

In addressing these challenges, the Bank of Italy can bring to the table its relevant experience, compared to other institutions, as it has already faced and solved many of the problems that the international statistical community is facing today. However, the scenario is constantly evolving and it requires continuous adaptation. In what follows we summarise the new frontiers which, in our opinion, should be taken in due consideration.

The challenge of data circulation in the international debate: access rights and data sharing

First, is the issue of the data circulation, both within a single authority (access rights) and between different authorities (data sharing). This is a recurring topic in the international debate. In this regard, the "European data strategy" recently launched by the European Commission plays a key role[38]. The goal is to "make the EU a leader in a data-based society" and create "a single data market", where data can circulate freely within the EU. In this context, the European Commission is evaluating a new European data governance model to facilitate data sharing among Member States. The data managed by central banks, despite being in most cases characterised by high levels of confidentiality and sensitivity profiles, and therefore subject to the obligation of professional secrecy or statistical secrecy, cannot escape these developments, at least with regard to the increase of circularity between authorities, thus avoiding double requests for data addressed to intermediaries.

[38] See European Commission (2020).

In the central banking community the debate on data circulation has been going on for some time[39]. Even the proponents of the strict separation between information used, respectively, for the supervisory and monetary policy functions, tend to recognise as virtuous policies of extensive data sharing between functions, such as, e.g., that adopted by the Bank of Italy for years. Among the advantages, it is emphasised the reduction of costs mainly borne by the reporting intermediaries, connected with the need to send the same data to two or more different functions or authorities (so-called double reporting). Where there is a policy of wide sharing of information, it is natural to make use of multi-purpose and integrated surveys, which optimise data collection and eliminate information redundancies.

The question unfolds along two main dimensions. The first dimension is linked to the growing importance of achieving a more efficient management of information by the authorities through economies of scale and scope and to the goal of bearing the minimum costs with the maximum benefits for citizens and businesses. The latter goal of avoiding unnecessary burdens is one of the inspiring principles of the so-called "better regulation"[40]. The second dimension is linked to the great importance of ensuring to the authorities, regardless of the specific institutional function assigned to them, access to a wide range of information, to allow their internal users a holistic view of the phenomena and to carry out analyses based on a more complete information set.

The sharing of data between European authorities[41] (data sharing), the removal of the legal and cultural constraints that still hinder it, the balance between the importance of sharing (the need to share referred to in the European data strategy) and the need to protect the confidentiality of information therefore constitutes challenging objectives for the coming years. The Bank of Italy contributes to this reflection in various statistical working groups, in particular within the Statistics Committee of the ESCB and the Committee on Monetary, Financial and Balance of Payments Statistics (CMFB)[42].

The challenge of integration in Europe

A second fundamental challenge concerns the finalisation of the project - which has been under discussion for some time - to implement the integrated approach described in the previous paragraphs for statistical, supervisory and resolution purposes. There is now a broad consensus on the advisability of proceeding towards the creation of a single European statistical dictionary to support the production of statistical, prudential and resolution reports. On this point, the considerations of the EBA discussion paper

[39] See: Draghi (2012), *"Speech at the Sixth ECB Statistics Conference: Statistics to deliver price stability and mitigate systemic risk",* 17 April 2012: "*Restrictive "blanket" confidentiality rules that cover all types of information, irrespective of whether there are real risks of a breach of confidentiality, are often counterproductive. They end up not only being damaging to information sharing and disclosure, but also to the real protection of confidentiality where and when this is needed. These rules confuse what is truly confidential and what needs strong protection with what is possibly market sensitive or what is simply inconvenient. So here is yet another area of activity on which statisticians and legal experts need to focus. They need to ensure that legal frameworks and existing practices find the right balance between the need for confidentiality and the possibility of data sharing".*

[40] See European Commission (2021), *Programme: Better lawmaking*. On 29 April 2021, the European Commission approved the communication "Better lawmaking", a tool to facilitate the European legislative process. This is a program aimed at legislative improvement in the European context, with the main objective of streamlining it to facilitate the recovery in Europe and adapt the laws to the future challenges. In this sense, it is essential to make the EU legislative process more transparent and open to the contributions of all interested parties and to encourage cooperation between the Member States of the Union, its institutions and civil society.

[41] See: *Irving Fisher Committee on Central Bank Statistics (2015).*

[42] The CMFB was established in 1991 with a European Council Decision (Decision 91/115 / EEC) with the aim of assisting the European Commission "in drawing up and implementing work programs concerning monetary, financial and balance of payments statistics". It is a Committee that brings together the experts of the European Statistical System (ESS) and the ESCB, which operates as a platform for dialogue and cooperation on the production of statistics belonging to these domains; moreover, since 2009, it has played an important role in the governance of the Excessive Debt Procedure.

on an integrated reporting system[43] and the arguments of the ECB in the document "The ESCB Input into the EBA feasibility report under article 430c of the Capital Requirements Regulation (CRR 2)"[44] are significant. Furthermore, on the basis of the experience of the Bank of Italy and other NCBs (including the Portuguese one), which over time have opted for the adoption of an integrated approach to data collection, the ECB has decided to launch a project for the cross-domain and cross-country integration and harmonisation of the information requirements envisaged in different statistical Regulations. At present, the IReF project exclusively concerns the data collected for statistical purposes but its success would be amplified if it were definitively associated, in the timing of implementation, with the complementary BIRD project. The latter is in essence, the European equivalent of the Italian PUMA cooperation initiative, illustrated in chap. 4.2. The joint implementation of IReF and BIRD would bring to Europe the countless advantages described above with reference to the integrated approach of the Bank of Italy.

In this scenario, the Bank of Italy will continue to actively contribute to the evolution of joint European decisions on the reporting obligations of the European banks. Consequently, at national level, it will be necessary to assess to what extent this evolution will impact on the reporting obligations imposed on intermediaries on the basis of national powers and to verify the possibility of their simplification.

The in-depth analysis conducted by the European Banking Authority on the integration of banking reporting

The current landscape of European banks' reporting obligations consists of different reporting regimes established mainly by prudential supervisors, resolution authorities and NCBs. Harmonised reporting obligations coexist with national reporting obligations. The current process of defining these obligations causes inefficiencies, as it does not provide for structured and systematic coordination mechanisms. On the contrary, an integrated European reporting system would create a single data ecosystem, with a significant reduction in the reporting burden of intermediaries, an increase in the efficiency levels of the processes on the side of the authorities, an easier sharing of information among the authorities, a better information governance.

These considerations are the basis of the reflections that the EBA is conducting on an integrated European banking reporting system, hopefully based on a single data dictionary capable of providing authorities and the banking industry with a coherent set of definitions, in line with the principle of non-redundancy ("define once").

The EBA discussion document on the feasibility study, under public consultation until 11 June 2021, provided an important opportunity to dialogue with all stakeholders on the future of the regulatory reporting system. The result of this debate will be reflected in the final feasibility study, required by Article 430c of the Capital Requirements Regulation (CRR II). On this occasion, the European

[43] See: EBA (2021).
[44] See: European Central Bank (2020).

Banking Federation (EBF), reiterated what it had long claimed regarding the current system of the reporting obligations of banks, both through an individual contribution and the publication of a joint communication with the European Association of Co-operative Banks (EACB), the European Association of Public Banks (EAPB) and the European Savings and Retail Banking Group (ESBG). In summary, the banking industry believes that an efficient solution should be built around the following principles: "*(1) Define Once: using a single data dictionary with all the data definitions; (2) Report Once: enhancing reusability and interoperability of the data; (3) Share Information: amongst the authorities instead of asking several times for data that has been already provided: (4) Enhanced Governance:* including the advice of the industry as the main stakeholder in every new data request ".

The challenge of addressing the new information gaps …

A third major challenge concerns the need to address new data gaps and to identify information no longer necessary in the reports currently acquired by the reporting agents. The information needs of the Bank of Italy, the ESCB, the SEVIF, the SSM and the SRM evolve over time; the data must be continuously enriched to meet the new needs of users. Furthermore, a continuous investment is essential to improve the quality of the data and expand their degree of detail. The reporting intermediaries, holders of basic information on their respective business, which is also constantly evolving, contribute decisively to the achievement of these general objectives.

… to be faced as part of a cost / benefit analysis

Evolutionary interventions in the reporting requirements cannot ignore the considerable costs incurred by intermediaries for the production of data. New requests from the authorities must therefore always be evaluated from a cost-benefit perspective. In the case of the Bank of Italy, these costs are reduced compared to those incurred by reporting agents in other jurisdictions thanks to the integrated solution adopted for statistical surveys. On the basis of the arguments reported in the previous paragraphs and of the experience earned up to now in our country, it can be stated that for banks and financial intermediaries the cost increase of additional information requirements is certainly lower in an integrated and multifunctional approach than in a silo approach. This is also thank to the support provided in the integrated system by the PUMA technical documentation.

In Italy, the cost-benefit analysis of the production process of credit and financial statistics has to also take into account the advantages deriving from the availability of feedback loops that the Bank of Italy elaborates by aggregating the report received and transmits to intermediaries. These are system-wide statistics, which each intermediary can use together with its own data in order to guide operational choices and shed light on its market positioning. Indeed feedback loops allow comparing each single intermediary’s position with the entire sector to which it belongs to or with groups of homogeneous operators, by size, branch of business, geographical distribution of operations.

Feedback loops for reporting intermediaries

Through the feedback loops, the Bank of Italy provides a set of statistical reports free of charge, both to the reporting intermediaries, i.e. the original producers of the data, and in some cases to the trade associations that represent them.

The feedback loops contain a wide range of information regarding many profiles of the intermediaries' business. The generation of flows is based on a specific set of metadata described in the statistical dictionary that is provided to intermediaries together with the data file, in order to make transparent the calculation processes of the indicators obtained from the elementary data.

At present the Bank of Italy produces numerous feedback loops that can be grouped into two broad categories, the statistical and the individual entity flows ; both categories are designed in close collaboration with the banking and financial industry so that their characteristics are such as to effectively meet the needs of the reporting entities.

The statistical flows contain aggregated data or other relevant indicators, which are derived from the elementary reports provided by the various intermediaries. Through these flows, reporting agents receive information that they can integrate into their internal information systems. The exploitation of these databases contributes to increase, both qualitatively and quantitatively, the management and control tools available to banks and financial companies as well as to reduce the cost/benefit ratio associated to the production of the original reports. In particular, each intermediary thus is provided with a unique information asset which, among other things, represents the basis for the calculation of market shares or other relative positioning and allows for timely monitoring of their evolution over time.

The individual entity flows, also called personalised feedback loops (e.g., those of the Central Credit Register, on analytical interest rates, on impaired loans, on historically expected losses, on default rates) are instead obtained by aggregating the information provided by various reporting agents on the same borrower/counterparty (natural or legal person) and are specific to each reporting agent. A particularly important flow concerns the data of the Central Credit Register. The Bank of Italy aggregates the data on the loans granted to the same counterpart by the various intermediaries and calculates the overall indebtedness of each counterpart towards the whole credit and financial system (the so-called global risk position); these flows are returned to intermediaries omitting the details on the individual financing intermediary.

Sharing this type of data with intermediaries pursues the objective of increasing the stability of the financial system, the degree of competitiveness and efficiency of the market and the sound and prudent management of supervised intermediaries. Indeed the availability of information on credit, including historical information, allows intermediaries to appreciate the evolution of the relationship with customers and contributes, together with a plurality of other information available to intermediaries, both to the credit granting decision process and the monitoring of the credit relationship.

The opportunity for periodic verification of the usefulness of the reporting obligations

However, the cost / benefit assessment should not only concern the phase of introduction of a new information requirement. The maintenance costs of such a complex and voluminous information framework are not negligible. It is therefore necessary to foresee periodic verifications of these costs following, e.g., the approach recently adopted by the EBA for the evaluation of the costs associated with the production of reports[45] or following the guidelines of the European Union aimed at introducing the "'One -In, One-Out "(OIOO) principle in case of new charges for European operators[46]. In line with this orientation, the Bank of Italy is strengthening a mechanism through which to request periodic confirmation from internal users of the effective usefulness of statistical data that have been stratified over time. This mechanism is useful for overcoming the physiological (and understandable) caution that users observe in consenting to the disposal of little-used portions of the information asset, even when they are very limited, in the hypothesis that these may regain relevance in a future moment. This approach, if not supported by in-depth assessments, can lead to redundancies and the proliferation of information, thus contrasting the culture, which is increasingly expanding, according to which essentiality in data is a driving force for efficiency.

Challenges in technology

A fourth challenge refers to the investments in technology to support the management of a mass of data that is growing at a rapid pace that is unprecedented in the past decades. In this context, it is necessary to invest in the large-scale adoption of innovative statistical methodologies for statistical production mentioned in par. 4. These methodologies, which the Bank of Italy has been committed for a long time, allow to increase the efficiency of processes and the quality of information.

The RegTech challenge

A fifth challenge that has emerged in recent years concerns the issue of RegTech, identified by the English regulator FCA (Financial Conduct Authority) as a subset of Fintech that focuses on the application of technology to make the provision of regulatory requirements by authorities more efficient and effective[47]. On this issue, it will be necessary to assess whether, in addition to the PUMA documentation - in Italy - and to the BIRD's - in Europe - it would be appropriate to hypothesise the drafting of the same reporting regulations in a form that allows for its automatic implementation by the intermediaries. This would contribute not only to facilitating compliance with regulatory requests but also to reducing the times necessary to obtain new structured and periodic information and to transform a new information requirement into a reporting obligation (on average it currently takes 2 years).

The PUMA documentation, which contains the transformation rules to generate the aggregate data, can be considered a sort of RegTech *ante litteram* (see Signorini, 2018), as it constitutes an important reference point for intermediaries in applying the reporting regulations. In spite of that the two solutions actually differs substantially. While both approaches require a standardised input data layer for all recipients of reporting obligations, in the case of PUMA documentation the timely production, consultation and use of this data layer are entirely voluntary. On the contrary, in a pure RegTech solution applied to the regulatory reporting, in which the reporting instructions are published in

[45] See *European Banking Authority* (2021a).

[46] See *Feasibility study: introducing "one-in-one-out" in the European Commission* (December 2019): "*Most recently, the new President of the European Commission, Ursula von der Leyen, announced that the Commission will apply the "'One-In, One-Out'(OIOO) principle "to cut red tape". In her mission letters to the designated members of the College of Commissioners, the new President stated that "the Commission will develop a new instrument to deliver on a 'One In, One Out' principle", adding that "every legislative proposal creating new burdens should relieve people and businesses of an equivalent existing burden at EU level in the same policy area"; and that the Commission "will also work with Member States to ensure that, when transposing EU legislation, they do not add unnecessary administrative burdens*". https://www.ceps.eu/wp-content/uploads/2019/12/Feasibility-Study.pdf.

[47] https://www.fca.org.uk/firms/innovation/regtech.

the form of an executable code, the production of the input data is regulated and therefore constitutes an obligation for the reporting agents.

RegTech applied to reporting regulations constitutes an innovative development that could add efficiency to the reporting systems and that is worth investigating. At present no convincing case studies or trials are available; it is therefore a new field to be explored in order to verify its practical applicability and suitability to speed up the availability of information.

The RegTech theme for intermediaries' reporting is interrelated with another important debate on two different and opposed approaches to data collection: the push and the pull approach. In the push approach the reporting agents send their data to the authority, while in the pull one it is the authority that has direct access to the intermediaries' information systems, in order to autonomously acquire the information it needs. In the abstract, the pull mechanism could have a number of benefits, both for reporting institutions and for authorities. For the former, it would reduce the certification processes of the data to be transmitted, as it would be only certified the compliance with the reporting regulations of the input data layer and not that of the individual reports. From the point of view of the authorities, they could activate the collection of information when they need it, improving the timeliness of access to information and saving on the costs of archiving a large amount of data. However, there are many concrete issues that still need to be investigated in order to move to a pull approach. At present the only application known to the authors is that of the Central Bank of Rwanda[48,] whose reporting system is, however, decidedly simpler than that of European countries.

The integrated exploitation of structured and unstructured data

Finally, a challenge particularly felt by users and which in recent years has also become a priority at an international level, concerns the need to jointly exploit the structured data collected from reporting entities and the unstructured information that is increasingly available to the community. We refer in particular to the data deriving from the ordinary activities that citizens, companies and institutions increasingly carry out in the digital world and that central banks can acquire from specialised providers, exchange them with other bodies and institutions or find them on the web[49]. Unstructured data are by their nature heterogeneous and fragmented and this makes their use complex, especially in conjunction with structured data.

The availability of these new data and their use in the analysis and decision making process, although still in an embryonic phase, first of all require the introduction of methodological and technological innovations (e.g., to create a data lake to be placed side by side with a data warehouse of traditional type). They also require carrying out an accurate verification of the choices made so far in terms of organisation and governance of the statistical function. That is, it will be necessary to evaluate whether to confirm the integrated approach with conviction, also considering the further facet that the new types of data assume, or to reformulate it to calibrate it on the new scenario. In fact the availability of these new data has a potentially disruptive impact also because it could induce a reduction in the reporting costs of intermediaries. Due to the intrinsic heterogeneity of the new types of data and to their multiple application areas, an agile and decentralised approach has been adopted in the Bank of Italy so far. According to it the various users have been able to autonomously experiment with the exploitation of this

[48] National Bank of Rwanda (2016 and 2017).

[49] These are digital traces of online purchases and searches, audio video and textual contributions on various social networks, data generated by digital devices. These data can be analised in a complementary way to the regulatory reports and / or used to anticipate the elaboration of some details of the official statistics.

type of information along a gradual approach aimed at satisfying individual information needs in the first place. The results of these experiments have converged into scientific works published in the Bank of Italy's research series and presented in numerous workshops and seminars.

On the horizon, therefore, the need for in-depth reflection arises on many issues. First of all, to decide the organizational profiles for the management of unstructured data. Secondly to attempt a classification or cataloging of this type of data, despite their unstructured nature, and establish minimum quality standards to allow their safe use. Thirdly, to facilitate their integration with traditional data originating from intermediaries. Lastly, to evaluate whether to govern the development of new machine learning methodologies and their application to new data and, if so, whether such governance should follow a single corporate strategy or is to be based on decentralised choices. These reflections will have to take into account that silo approaches are considered nowadays counterproductive. Moreover, it is not possible to ignore the widely shared reflections underway in Europe on the opportunity of joint purchases of information that the large commercial data providers own. In the background there is the main objective that has long been inspiring the approach followed by the Bank of Italy in information management: adopting a data-driven approach, efficiently exploiting the rich supply of data in decision-making and research processes.

The scenario is evolving, but the main drivers are still very topical

In this new context, most of the fundamental issues addressed in this paper continue to play a central role, albeit declined in a new way. Notably: (a) the goal of integration, albeit between different types of data - structured and unstructured - and information sources; (b) the central role of the statistical dictionary, to cover all types of information supporting decision-making processes and for the benefit of users; (c) the unitary governance of the various and more diversified information segments - also in terms of redefining access profiles; (d) the quality standards to be ensured for the new types of data.

Everything illustrated so far insists on the idea that information asset must be considered as a strategic resource for the benefit of all the institutional functions of a central bank, that nowadays operates within a much broader and increasingly interconnected environment than in the past. In this new context, data governance - internal and coordinated with other central banks - must be adequate to extract value in the most possible efficient way from the large amount of available data.

Although, as we have seen, the changes in the scenario are tumultuous, the objectives set by the Bank of Italy also in the relationship with the other partners continue to be fundamentally the same. In short, it is necessary to ensure that there is clarity on the meaning of each data, on the allocation of responsibilities between users and IT staff, on criteria for establishing the adequacy of the quality of a data, on the protection of confidentiality of information. Ultimately it is necessary to ensure that a common vision on all these issues is shared within the whole organization. In the background, the main objective that has long inspired the approach followed by the Bank of Italy in the management of credit and financial information remains that of allowing users - internal and external - to efficiently exploit the very rich asset of data for the decision-making processes and economic and financial research.

Final remarks

The Bank of Italy collects a wide range of data from banks and financial intermediaries. The data are then subject to a highly detailed data quality and compilation process which culminates in the release of many statistics relevant to internal and external users. To properly manage and support such a detailed process, an integrated approach is adopted.

The process is structured in several sequential phases: (a) detection and preliminary analysis of new information needs; (b) development of the information project, which culminates in the design of the data collection. In this phase, different information needs are put together in order to find, whenever possible, the data level of detail that satisfies all purposes; (c) direct dialogue with reporting agents, which mainly takes place within the PUMA cooperation initiative; (d) drafting the PUMA documentation, that supports reporting entities in identifying relevant raw information in their internal databases and in producing the required aggregates; (e) data acquisition and validation by the Bank of Italy; and (f) dissemination of the validated data to the internal DWH and to some qualified external parties.

The data management process involves a large number of entities other than the Bank of Italy and the reporting intermediaries themselves. These are IT companies that supply software for reporting, service companies offering data management services and other authorities defining their own reporting requirements. The work also requires the deployment of various skills from: statisticians, data administrators, IT developers, data analysts, economists and researchers. The basic organisational solutions for this process have been designed with a view to obtaining sound overall efficiency and effectiveness. The data model is represented by a general matrix that accommodates all data needs regardless of their ultimate use and the information is described by a single statistical dictionary that ensures the uniqueness of each concept. Backing all these steps is the on-going cooperation between the Bank and the reporting intermediaries, since the latter are the primary producers of the information. The entire process is governed by a collegiate body: the Bank of Italy's Statistics Committee. Following a unitary and comprehensive approach, the operational data management is assigned to a specific organizational unit, the Statistical Data Collection and Processing Directorate.

The paper details the various facets of the integrated approach that the Bank of Italy has applied since the late 1980s for the management and governance of credit and financial data collected from intermediaries. This is a framework whereby information is managed according to a holistic and multi-purpose vision that removes, right at the outset, the possibility of creating information silos and it does not allow redundancies. A corollary of this comprehensive approach is the establishment of a broad information circulation policy within the Bank of Italy so that access to the data is authorised on a need-to-know basis and then, exceptions apart, users are guaranteed access to the full DWH.

It is important to note that when the integrated approach was first designed and implemented, there were basically no other central banks using the same methodology and philosophy. Over the years, however, this strategic choice has proved to be far-sighted and it has been recognised as a success in addressing the growing demand for the structured data produced by intermediaries.

For many years, the Bank of Italy has been promoting its approach in the European fora in which it participates, above all in the ECB Statistics Committee. Over the last decade it has been wholeheartedly embraced by the ESCB, resulting in the launch of the following ESCB projects: IReF, BIRD, RIAD and CSDB. The first two are still to be finalised, whereas the others are now well consolidated. An integrated view of statistical reporting also inspired the decision of the European legislator to entrust the EBA with the task of preparing a feasibility study on the integration of all reporting obligations weighing on European banks. The expectation, especially on the part of banks, is that the rationalisation and simplification of reporting obligations will be achieved within the next five years and, as a consequence, the overall efficiency of the production processes for ESCB statistics and supervisory and resolution reports will increase.

However, the evolution of the external environment - in terms of phenomena to be monitored, types of available data, level of detail requested by users - nowadays requires NCBs to face new challenges that, in turn, have a significant impact on data management and governance choices. Nevertheless, it should be noted that, also in the past, many important changes heavily affected reporting obligations.

Today, however, the urgency, complexity and magnitude of the challenges may even call into question the choices made over time concerning the methodology adopted for data management. Authorities are aware of the need to establish a new form of global information governance that takes into account the growing supranational dimension of regulatory reporting and the growing importance of non-conventional data. The goal is to allow for an increasingly effective management of the information used to fulfil the authorities' respective mandates and to exploit the new potential that technology offers to extract rapidly the essential and accurate information inputs from large volumes of diverse (and often unstructured) data. In fact, one of the currently most demanding challenges is how to enable the joint use of data collected by intermediaries together with the enormous amount of unstructured information offered by the digital societyUnstructured data are by their very nature very fragmented. From a methodological and technological point of view, this fragmentation makes it very complex to make use of them and, above all, to do so together with structured data. Besides resorting to artificial intelligence methods, the technological solution to this problem seems to be the development of a data lake (or data mesh), no matter how challenging this may be. In this regard, the Bank of Italy will have to evaluate the most suitable organisational model - be that centralised or decentralised - to manage this new category of data, keeping in mind that the disadvantages of information silos most likely also apply to non-conventional data. Lastly, it also appears necessary to establish fruitful cooperation systems among the various authorities (national and international) that are engaged in statistical production, in order to pool their economic and organisational efforts when acquiring this new type of information. In this regard, important impetus is provided by the EU data strategy.

In this rapidly evolving context, the ESCB organizational structures responsible for the management of statistical information have already started to review their systems. In fact, they have planned (and in some cases already implemented) important organisational reforms and introduced new governance roles to ensure the sound management of the corporate information assets. Such initiatives also aim at ensuring the most widespread synergy with IT and at improving the quality of the service offered to internal and external users. This brainstorming of ideas and solutions is enhanced by the continuous dialogue among national and supranational authorities, which interact on these issues at different negotiating tables. With particular reference to the ESCB, NCBs and the ECB are working together to make the European statistical framework more efficient overall, thus looking beyond the national borders. The concrete results of these efforts are the IReF and BIRD projects.

In such a stimulating scenario, which has the potential of setting new foundations for the work of the NCBs' statistical departments in the coming years, reporting intermediaries will continue to play a key role in the production of the necessary information for NCBs and supervisory authorities. A careful plan for the new reporting obligations is highly advisable so as not to have to act quickly in the face of problems as they arise. Last-minute solutions, in fact, hinder the possibility of establishing a fully efficient reporting framework because, in the end, the data collection turns out to be a series of subsequently added layers of information that are then rather difficult to streamline. For the smooth and effective development and production of information in Europe, we therefore do not see any alternative other than the establishment of coordinated and appropriately planned action by all the relevant authorities.

The Bank of Italy is an important player in this scenario. By virtue of the many years of experience gained in the field of credit and financial information collection, the Bank is a respected counterparty in all international fora. It continues to offer the ESCB essential support in its effort to make data collection and processing more efficient, identify new and more appropriate governance solutions, and enhance the dialogue and collaboration with both intermediaries and data users.

Glossary

Banks' Integrated Reporting Dictionary (BIRD): it is an ESCB project that aims to support banks to organize, in the most efficient and effective way, the information stored in their internal IT systems in order to produce the reports required by European Regulations. BIRD represents an integrated and standardised dictionary including a detailed description of the information to extract from the banks' internal databases and the transformations that are needed to produce the regulatory reporting at European level. BIRD's dictionary is developed and maintained in close cooperation between the ESCB and the banks that participate in the initiative on a voluntary basis; it is disseminated free of charge on the ECB's website. Its adoption in the reporting system is optional.

Centralised Securities Database (CSDB): securities repository managed by the ECB, which contains detailed and up-to-date information on all individual securities of interest to the ESCB: securities issued by euro area residents, transacted between euro area resident or denominated in euros, regardless of the issuer's residence.

Classification variable: variable that characterises specific attributes of an economic phenomenon or entity such as, e.g., residence, country and institutional sector of the counterparty, currency of denomination of the amounts, duration of the transaction.

Data collections on individual entities (or Nominative surveys): they are the regulatory reports where information refers to an individual counterpart entity, be that a natural or legal person (e.g., the debtors and guarantors of the Central Credit Register, the issuers of financial instruments, counterparties of economic transactions, members of the corporate bodies of supervised intermediaries, public administration bodies). These data collections are based on the use of the unique entity identifier (subject's code) which is managed in the Entities register.

Data lake: it is a store of data including raw information, sensor data, social data and transformed data. It supports the storage of information when the following factors become critical: i) the data volume is particularly high; ii) the data are characterised by a wide variety of formats and are typically unstructured; iii) the data to be processed and their representation methods change very frequently.

Domain in use: it is the subset of the elements of a domain of values that a specific phenomenon or classification variable can actually take.

Domain of values: it is the complete set of possible values that can be taken by a classification variable or by an economic phenomenon observed in a statistical unit of a population or by a subset of statistical units selected on the basis of a given criterion.

Entities register: it is the archive containing reference data on individuals (i.e. natural persons, companies, public administration bodies, other entities). It is used by the Bank of Italy to support the national data collections of quantitative information on individual entities and it is also used to collect granular ESCB statistics (AnaCredit). Each subject is assigned a unique code that allows the exact identification. Its importance has grown over the years, together with the increase of reporting of individual entities data and the wider use of granular data in economic and macro-prudential analyses. The Bank of Italy's register also plays an important role within the ESCB, as it is one sources for the Register of Institutions and Affiliates Data (RIAD) managed by the ECB.

Each entity is described by a series of reference data (i.e descriptive attributes) which mainly depend on the type of subject (natural or legal person)[50]. The attributes allow the transformation of granular data into aggregated statistics. The content and structure of the register is independent from the peculiarities of the various data collections on individual entities that use it. The Italian Entities register is fed with official sources (e.g. public registers, official lists and registers managed by institutions that certify the existence of the entities and the validity of their personal data) or with the input cooperatively provided by the reporting entities or other reliable sources. A copy of the register is shared with the banking system in order to support the production of the reporting obligations.

Information segment: the term is used to refer to the economic and business areas studied and monitored by the Bank of Italy in relation to its functions (e.g. monetary and financial statistics, statistics on payment systems, prudential information).

Integrated approach: in the note, this approach refers to the model adopted by the Bank of Italy since the late 1980s to produce a considerable part of its information asset, i.e. the part based on reports provided by banks and other financial intermediaries. The core of the integrated approach is that it is (almost) always possible to put together the different users requirements and drill-down to identify data points that can be used by multiple users; moreover, each data point must be collected only once. On these grounds, the management and governance of information can be achieved with a centralized framework and common methodologies, infrastructures and tools. The integrated approach makes use of the relationships and synergies existing between the different parts of the information system. This allows avoiding data duplication, thus helping to minimise the burden on reporting entities. The integrated approach enhances the multi-purpose use of a single data and the holistic vision of the corporate information-statistical system. It is the opposite of the so-called "silos approach" whereby the information system is split into separate portions in which information, processes and management platforms are not shared. This approach usually involves redundancy of data and information and it is subject to duplication of processes and infrastructures.

The integrated approach, adopted by the Bank of Italy to produce credit and financial statistics, is based on the following pillars:

- a unitary process to detect the information needed by the institutional functions from bank and financial intermediaries. The process requires to establish a high degree of coordination by the organizational unit in charge of the management the system;
- the application of integrated and granular schemes to design the data collections that simultaneously satisfy different user needs. This is achieved by choosing the appropriate level of granularity, detail and frequency of data collection;
- a single statistical dictionary with complete and non-redundant descriptions of all the statistical concepts and metadata belonging to the information system, including relationships among variables and transformations to obtain derived statistics and indicators;
- a single statistical data warehouse, accessible by all internal users on the basis of well-defined access rules;
- a single and dedicated IT platform to collect, manage and disseminate all the statistical data;

[50] See Circular of Bank of Italy n. 302 – June 2018.

- the concentration of the data quality process in one organisational Unit (the Statistical Data Collection and Processing Directorate), which manages the various components of Bank of Italy's information asset with a comprehensive approach;
- the coordination with reporting intermediaries of the activities in order to define uniform rules to extract the raw data from their internal information systems and to produce the regulatory reports. This is done in the PUMA working groups.

Integrated Reporting Framework (IReF): it is an ESCB project that seeks to integrate the existing ESCB statistical data requirements for banks, as far as possible, into a unique and standardised reporting framework which is harmonised among statistical domains and across countries. At the beginning the IReF will primarily focus on ECB requirements relating to banks' balance sheet data, interest rates statistics, securities holdings statistics and (granular) credit data. Furthermore, to maximise the benefits of the integration, the IReF will also take into account the requirements arising from national needs (as long as they are shared among NCBs) such as, for example, those for the compilation of the balance of payments and financial accounts. In the initial phase, the IReF will exclude data that do not directly relate to banks' balance sheet assets and liabilities, such as the ECB payment systems or money market statistics requirements.

Matrix model: it is the data representation model used in the Bank of Italy's integrated information system. The information is described with a double entry table (the matrix), where the rows host the phenomena and the columns show the characteristics that qualify the phenomena. For each cell of the matrix, different measures of a phenomenon can be represented. This model is well suited to represent very articulated phenomena, as it allows to normalise the representation of the information as coordinates of a matrix, regardless its specific use.

Merits and costs procedure (M&C): it is a process, required by Regulation (EC) 2533/98, to ensure that new statistics are sufficiently justified by high-priority policy needs; the process provides incentives to search for the most cost-effective solutions, with the aim to keep the burden placed on the reporting agents to a minimum. A comprehensive description of the M&C procedure is available on the ECB web site[51].

Primary reporting and Secondary reporting: primary reporting is the term used to define the regulatory reports that national reporting agents send to the NCB, whereas secondary reporting are the data flows, based on same data (or their aggregations, as applicable), which NCBs send to other authority. The latter is typically the ECB but it can also be the EBA, the Bank for International Settlements, the International Monetary Fund, etc. Secondary reporting is intended to cover information needs ruled by Regulations, Guidelines or other legal acts issued by the receiving authority. Therefore, also from a methodological point of view, a secondary reporting is defined according to contents, data representation models, formats and coding systems set by the recipient authority. Primary and secondary reporting are two fundamental phases of the data collection and dissemination process.

[51] https://www.ecb.europa.eu/stats/ecb_statistics/governance_and_quality_framework/html/merits_costs_procedure.en.html

Register of Institutions and Affiliates Data (RIAD)[52]**:** it is the ESCB database of reference data on legal and other statistical institutional units, the collection of which supports business processes across the Eurosystem and the performance of the tasks of the ESCB and the SSM. It is disciplined by ECB Guideline 2018/16. RIAD contains a wide range of attributes on individual entities and relationships between such entities that enables the derivation of group structures. The register is used to prepare the official lists of monetary financial institutions, investment funds, financial vehicle corporations engaged in securitisation transactions, payment statistics relevant institutions and insurance corporations. It is also the reference register for the reporting of AnaCredit granular data to the ECB.

RegTech: it is, in broad terms, the use companies make of IT tools to support and verify the compliance with rules, regulations, laws and reporting requirements. RegTech solutions appear to be growing especially in highly regulated sectors, such as banking and finance. Indeed, over the last few years in these sectors it increased the need to better manage, thanks to the use of new technologies, a growing number of legislative and regulatory obligations that are particularly demanding in terms of the time taken and which require the analysis of large amounts of information. From the point of view of supervisory authorities, RegTech can be an important tool to verify regulatory compliance, as RegTech solutions aim at facilitating flows of information between regulated companies and regulators and providing the latter with data they need most. With regard to reporting obligations, RegTech responds both to the need of banking and financial institutions to smoothly producing reports in line with regulatory requirements, and to that of reporting entities and regulators to define collaborative reference systems that allow reducing the time needed to produce and verify data. The range of RegTech instruments is wide and in evolution. By way of example, the followings can be mentioned: the use of cloud computing solutions, which allow a remote data management in a secure and flexible way, or the adoption of artificial intelligence techniques, to improve the organization in the archives, the efficiency of data extraction, the automation of the regulatory reporting or the exploration of large masses of data.

Securities register: it is the Bank of Italy's database containing the details of financial instruments issued or transacted by residents in Italy and that banks, financial intermediaries and other supervised companies include in their regulatory reports for the Bank of Italy. It also contains information related to the issuers. The securities are uniquely identified by the International Securities Identification Number (ISIN), which is assigned by the Bank of Italy in its role of National Numbering Agency for Italy. A monthly report of the Securities register is provided to intermediaries to allow them to comply with their reporting obligations. The securities register also represents the source of the Italian contribution to the ESCB Centralised Securities Database.

Semantic integration: it is the process aimed at conceptually linking information from different sources in order to ensure homogeneity of definitions and taxonomy. The availability of a single statistical dictionary is fundamental to achieve a proper semantic integration.

Statistical data warehouse (DWH): it is the environment for the storage of all the elementary structured data (collected from the reporting entities or acquired from external suppliers) and of all the indicators and aggregated statistics calculated on the elementary data. The DWH information is made available to all internal users who request it, according to access rules established by the Bank

[52] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018O0016&from=EN

of Italy's Statistics Committee. In the DWH there are no rigid separations between information segments; rather, the access rules are based on the actual users' needs for the performance of their tasks (the so-called need to know principle). Apart from some exceptions, access rights are not differentiated on the basis of the information segment or institutional function (e.g. prudential analysis, economic research, statistical production or markets supervision and payment systems oversight); at the same time it is possible to adopt criteria allowing to protect the confidentiality of individual data. The DWH is at the heart of the integration of the information system, as it allows the use of data originating from multiple sources by users belonging to the various institutional functions of the Bank; this is possible thanks to the recourse to applications allowing data navigation and to the use ofa single statistical dictionary.

Statistical dictionary (or data dictionary): the set of all the metadata that allow to describe in a formal and standardised language and in a non-redundant way the information system data, the rules for transforming them into aggregates or indicators, the data quality validation rules, the structure of reports and that of databases in which they are stored. It includes, e.g., the list of the phenomena's classification variables, the set of values' domains and information on the unit of measurement. In the Bank of Italy the statistical dictionary constitutes a pillar of the integrated system for the production of credit and financial statistics, as it represents the substrate that allows linking the various phenomena reported in the system without redundancy (so-called semantic integration). It can be considered as a sort of common language to describe the phenomena themselves and documents elementary data transformations into aggregated statistics.

Structured data: information organized according to a predefined representation model (the data model) and adhering to a predetermined set of rules that define its type, attributes (e.g., reference date, statistical unit identification code, geographical location, type of amount, etc.) and relationships. The data model is organised in rows and columns; they are typically stored in a relational database, from which they can be retrieved, separately or in a variety of combinations, for processing and analysis purposes.

Unstructured data: data that do not adhere to a predefined representation model . Examples of unstructured data are text, images, audio and video, e-mails, spreadsheets and other objects stored as files or quantitative information that is not organised according to pre-defined rules. Unstructured data can have very different origins (extracted from human language, acquired through sensors, extracted from social media); moreover, they tend to occupy volumes much higher than those of structured data. For these reasons, the IT tools and data administration techniques typically applied for structured data are not applicable. In recent years the development of new tools to extract, store (via data lake) and manage this type of information appear to be growing. Among the new tools there are: data mining methods to perform large-scale processing, artificial intelligence techniques, natural language processing methodologies to assign meaning to business documents, emails, magazine articles and social media posts, pattern recognition algorithms to identify people, animals or other objects in digital images and videos, text-to-speech conversion to convert audio videos in searchable text. These tools were accompanied by an intense use of innovative and increasingly sophisticated statistical techniques (machine learning) to facilitate their exploitation.

References

Bank of Italy (2020), *The Central Credit Register*, The Bank of Italy Guides, October (https://www.bancaditalia.it/pubblicazioni/guide-bi/guida-centrale/index.html?com.dotmarketing.htmlpage.language=1)

Basel Committee on Banking Supervision (2020), *Progress in adopting the Principles for effective risk data aggregation and risk reporting*, April (https://www.bis.org/bcbs/publ/d501.htm)

Bernardini M., Massaro P., Pepe F. and Tocco F. (2021), *The market notices published by the Italian Stock Exchange: a machine learning approach for the selection of the relevant ones,* Occasional Papers, n. 632, July (https://www.bancaditalia.it/pubblicazioni/qef/2021-0632/QEF_632_21.pdf)

Casa M., D'Alessio G. (2020), *Le statistiche della Banca d'Italia nell'epoca del coronavirus*, Bank of Italy, Note Covid-19, October (https://www.bancaditalia.it/pubblicazioni/note-covid-19/2020/Nota-COVID-Statistiche-2020.10.22.pdf)

Cusano F., Marinelli G., Piermattei S. (2021), *Learning from revisions: a tool for detecting potential errors in banks' balance sheet data,* Bank of Italy, Occasinal Papers, n. 611, March (https://www.bancaditalia.it/pubblicazioni/qef/2021-0611/QEF_611_21.pdf)

Del Vecchio V. (2007), *The "Matrix" model – Unified model for statistical data representation and processing*, May, (https://www.bancaditalia.it/statistiche/raccolta-dati/sistema-informativo-statistico/modellazione/matrixmod.pdf)

Draghi M. (2012), *Speech at the Sixth ECB Statistics Conference: Statistics to deliver price stability and mitigate systemic risk,* April (cb.europa.eu/press/key/date/2012/html/sp120417.en.html)

Draghi M. (2016), *Welcome address at the Eight ECB Statistics Conference "Central bank statistics: moving beyond the aggregates"*, Frankfurt am Main, 6 July

European Banking Authority (2021), *Discussion paper on integrated reporting*, March, (https://www.eba.europa.eu/eba-launches-discussion-paper-integrated-reporting)

European Banking Authority (2021), *Discussion paper on a feasibility study under article 430c of the Capital Requirements Regulation (CRR 2),* March (https://www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Discussions/2021/Discussion%20on%20a%20Feasibility%20Study%20of%20an%20Integrated%20Reporting%20System%20under%20Article%20430c%20CRR/963863/Discussion%20Paper%20on%20integrated%20reporting.pdf)

European Banking Authority (2021), *Study of the cost compliance with supervisory reporting requirements* (ba.europa.eu/sites/default/documents/files/document_library/Publications/Reports/2021/1013948/Study%20of%20the%20cost%20of%20compliance%20with%20supervisory%20reporting%20requirement.pdf)

European Banking Federation (2021), *EBF response to the EBA discussion paper on the feasibility study of an integrated reporting system under article 430c CRR*, June (https://www.ebf.eu/wp-content/uploads/2021/06/EBF_045187-EBF-response-to-EBA-consultation-on-Integrated-Reporting-Key-Points.pdf)

European Central Bank (2010), *The Centralised Securities Database in brief*, February (https://www.ecb.europa.eu/pub/pdf/other/centralisedsecuritiesdatabase201002en.pdf).

European Central Bank (2016) – *ECB banking supervision: SSM priorities 2016*, January (https://www.bankingsupervision.europa.eu/ecb/pub/pdf/publication_supervisory_priorities_2016.en.pdf?024a0072fe923441556e5bba7251dd6d)

European Central Bank (2016), *The ECB's merits and costs procedure in the field of European statistics*, July (https://www.ecb.europa.eu/stats/ecb_statistics/governance_and_quality_framework/html/merits_costs_procedure.en.html)

European Central Bank (2018), *SSM Supervisory Manual – European banking supervision: Functioning of the SSM and supervisory approach*, March (https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisorymanual201803.en.pdf?1584b27046baf1e68f92f82caadb3a63)

European Central Bank (2018), *Guidance notes to reporting agents on SHS regulation for statistics on holdings of securities by reporting banking groups (SHSG),* September (https://www.ecb.europa.eu/pub/pdf/other/guidance_notes_to_reporting_agents_on_shs_regulation201809.en.pdf)

European Central Bank (2020), *The ESCB input into the EBA feasibility report under article 430c of the Capital Requirements Regulation (CRR2)*, September (https://www.ecb.europa.eu/pub/pdf/other/ecb.escbinputintoebafeasibilityreport092020~eac9cf6102.en.pdf).

European Central Bank (2020), *The Eurosystem Integrated Reporting Framewrk, an overview*, November (https://www.ecb.europa.eu/pub/pdf/other/ecb.escbirefoverview202011~ebb404b7b6.en.pdf).

European Commission (2020), A *European strategy for data*, (https://digital-strategy.ec.europa.eu/en/policies/strategy-data and https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2102)

European Commission (2020) *Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data governance Act)*, November (https://digital-strategy.ec.europa.eu/en/policies/data-governance)

European Commission (2021), *Better regulation: why and how,* April (https://ec.europa.eu/info/law/law-making-process/planning-and-proposing-law/better-regulation-why-and-how_it).

Eurostat (2017), *Codice delle statistiche europee adottato dal Comitato del Sistema Statistico Europeo*, November (https://ec.europa.eu/eurostat/documents/4031688/9394142/KS-02-18-142-IT-N.pdf/2d3874da-4253-4f20-9cfd-304f48a5ed1a)

Financial Stability Board, International Monetary Fund (2009), *The Financial crisis and information gaps - Report to the G-20 Finance Ministers and Central Bank Governors*, October (https://www.imf.org/external/np/g20/pdf/102909.pdf)

Financial Stability Board, International Monetary Fund (2010) *The Financial crisis and information gaps - Progress Report Action Plans and Timetables,* May (https://www.imf.org/external/np/g20/pdf/053110.pdf)..

Financial Stability Board (2011), *Shadow Banking: Strengthening Oversight and Regulation - Recommendations of the Financial Stability Board*, October (https://www.fsb.org/wp-content/uploads/r_111027a.pdf)

Giudice O., Massaro P., Vannini I. (2020), *Institutional sector classifier: a machine learning approach*, Bank of Italy, Occasional Papers, n. 548, March (https://www.bancaditalia.it/pubblicazioni/qef/2020-0548/QEF_548_20.pdf).

International Monetary Fund (2003), *Data Quality Assessment Framework*, February (https://www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm)

International Monetary Fund (2007), *Italy - Report on the Observance of Standards and Codes – data Module*, February (https://www.imf.org/en/Publications/CR/Issues/2016/12/31/Italy-Report-on-the-Observance-of-Standards-and-Codes-ROSC-Data-Module-20482)

Irving Fisher Committee on Central Bank Statistics (2015)*, Data-sharing: issues and good practices - Report to BIS Governors prepared by the Task Force on Data Sharing,* January (https://www.bis.org/ifc/events/7ifc-tf-report-datasharing.pdf)

Lautenschlager S. (2018), *20 years of ESCB statistics: Past achievements and future challenges,* Speech at the Ninth ECB Statistics Conference "20 years of ESCB statistics: What's next", Frankfurt am Main, July

National Bank of Rwanda (2016), *Annual report* (https://www.bnr.rw/news-publications/publications/annual-reports/).

National Bank of Rwanda (2017), *Annual report* (https://www.bnr.rw/news-publications/publications/annual-reports/).

O'Leary Daniel E. (2013), "*Big data, the Internet of things and the Internet of signs*" (https://onlinelibrary.wiley.com/doi/full/10.1002/isaf.1336)

Padoa Schioppa T. (1993), *Le segnalazioni statistiche rivolte alla Banca d'Italia come fondamento della gestione delle aziende di credito*, *("The statistical reporting for the Bank of Italy as a fundamental pillar for business management in banks"),* speech at the first convention on "Information as a strategic management resource"- Rome, 12 February 1993, organised by Associazione Italiana per la Pianificazione ed il Controllo di Gestione in Banca e nelle Istituzioni Finanziarie (the Italian association for planning and management control in banks and financial institutions).

Santomartino A. (2012), *La produzione statistica della Banca d'Italia, l'utilizzo da parte degli intermediari bancari,* Bancaria, n. 3, March (https://bancaria.it/livello-2/archivio-sommari/gli-ultimi-sommari-di-bancaria/bancaria-marzo-2012/la-produzione-statistica-della-banca-d-italia-l-utilizzo-da-parte-degli-intermediari-bancari/)

Signorini, L. F. (2012), *Le banche italiane verso Basilea 3*, speech at the convention "ABI Basilea 3", Rome, 26 June (https://www.bancaditalia.it/pubblicazioni/interventi-vari/int-var-2012/signorini-260612.pdf).

Signorini, L. F. (2018), Introductory speech at the workshop *"Tra segnalazioni nazionali e reporting armonizzato europeo: rafforzare la cooperazione tra gli intermediari e le autorità" (*Between national and harmonised reporting in Europe: strengthening the cooperation between reporting agents and authorities*),* May (https://www.bancaditalia.it/pubblicazioni/interventi-direttorio/int-dir-2018/Signorini-15052018.pdf)

UNECE (2009), *Generic Statistical Business Process Model*, April (https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.61/2009/mtg1/zip.32.e.pdf).

United Nations (2014) "Fundamental Principles of Official Statistics", March, (https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf)

Zambuto F., Buzzi M. R., Costanzo G., Di Lucido M., La Ganga B., Maddaloni P., Papale F, Svezia E. (2020), *Quality checks on granular banking data: an experimental approach based on machine learning*, Banca d'Italia, Occasional Papers, n. 547, March (https://www.bancaditalia.it/pubblicazioni/qef/2020-0547/QEF_547_20.pdf).

Zambuto F., Arcuti S., Sabatini R., Zambuto D. (2021), *Application of classification algorithms for the assessment of confirmation to quality remarks,* Occasional papers, n. 631, July (https://www.bancaditalia.it/pubblicazioni/qef/2021-0631/QEF_631_21.pdf)