Data Fitness for Integration

Bernadette Lauro (bernadette.lauro@ecb.int)[1] Raffaella Traverso (raffaella.traverso@ecb.int)[1]

Keywords: data integration, master data, data quality

1. INTRODUCTION

Data are at the heart of policy decisions and represent the most valuable asset, after people, at the European Central Bank (ECB).[2] This recognition has driven the institution towards the adoption of a data management strategy which is oriented towards more common and integrated data processes and to a data centric architecture.

Data quality management is core to the ECB statistical function. In the last two decades *tremendous* efforts have been undertaken in order to fill the gaps in aggregated and standardised data. Today the focus shifted *on managing high volumes of data, in producing high-quality and granular data*[3] and in sharing data products. This has driven the ECB towards the adoption of large-scale IT systems that allow to combine more efficiently data sources and data models and to perform data analysis with different analytical tools. However, combining data in an IT platform, although necessary, is not sufficient to achieve high quality integrated data.[4]

Indeed, the quality of integrated data depends not only on the quality of the individual data sources, but also on the quality of all interrelated components of data management. These components encompass the univocal identification of business entities (master data) for which economic transactions and positions need to be analysed. Further components are the conceptual definitions and the methodology of compilation of data collected and provided from different sources; as a consequence, contextual integration requires congruence in the definition of concepts and in the codification of data. Additionally, clear and structured information (metadata) that clarify the meaning and the structure of the data are relevant to integrate data efficiently. Finally, a state-of-the-art technical infrastructure is essential to enable the integration of all these components.

Integrating data involves different but interdependent data management activities. Therefore, to reach the desired level of quality, not only data but also models, processes and tools must respond to measurable quality indicators, according to a defined and compatible maturity model. In this sense, the fitness for purpose of integrated data is a richer concept than the fitness of a single dataset individually considered.

2. THE CHALLENGES OF DATA INTEGRATION: A USE CASE

Data analysis is a fundamental activity supporting the monetary-policy and supervisory functions at the ECB. The increased need to analyse granular data on individual assets of

[1] European Central Bank.

[2] "Data are not simply addenda or second-order artifacts; rather, they are the heart of much of the narrative literature, the protean stuff that allows for inference, interpretation, theory building, innovation, and invention" (Cronin, 2013, p. 435)

[3] Cœuré (2017)

[4] By data integration we intend the combination of data residing in different data sources with the scope of providing a unique view of these data

banks has stimulated scattered-local data management activities.[5] Yet, providing "fast, easy to access, integrated and high quality data", that is data at users' fingertips, has become a priority for the overall ECB data management strategy.

In order to attain this vision, data management activities should pursue criteria of efficiency, effectiveness and preservation of data confidentiality. Data quality, defined as "fitness for purpose", measures the achievement of this goal. Hence, we pose the question: how is the quality of source data affected when data are integrated?

To answer this question, we present the challenges posed by data integration by way of a use case: the integration of data on banks' securities.

The availability of different sources for granular data on securities – both held and issued by individual banks - has engaged ECB analysts into ad-hoc data integration activities. This has motivated the initiation of a pilot on data integration to obtain the funding and assets of banks (FAB). In this project data from commercial datasets are combined with institutional data and enriched with information on securities prices and ratings. Quality checks are performed to ensure that trustworthy information underlie reports supporting the decision making process. The final aim is to prepare tailored and agile reports based on granular data. Given the integrated output, it's possible to prepare benchmarks with other datasets and aggregation tools for data users. The output datasets on bank groups' securities information are provided on a common IT infrastructure that allows for storing large-scale data (the Data Intelligence Service Centre - DISC platform).

Integrated products are useful for many tasks at the ECB: the analysis of how the ECB's Asset Purchase Programme influences banks' portfolio rebalancing or about financial risks embedded in securities holdings.

3. DATA MANAGEMENT ACTIVITIES INVOLVED IN THE FAB PROJECT

The integration of granular data encounters a number of issues, each related to specific data management activities. We group these activities in four main domains of integration, namely: a) master data, b) semantic c) metadata and d) technical tools integrating activities.

3.1. Master data integration

The first step in the FAB project is to identify univocally the bank – our entity of interest - for which we want to analyse the securities business. For this, we avail ourselves of master data, that is, data that identify the bank in the most accurate way, and that therefore are considered the authoritative source for the identification of the business entity. Integrating master data consists in consolidating different reference data sources. Only after this exercise it is possible to have a view of the securities activities in the balance sheet of the specific bank. In simplified terms, in a reference data source a specific bank ID is associated to its name, address, hierarchical structure and legal form. However, as shown in **Table 1**.**A** different identifiers might be associated to the same

[5] Granular data collected by the ECB include data on security-by-security information contained in the Centralised Securities Database (CSDB), loan-by-loan in the Analytical Credit datasets (Anacredit) and holdings of security assets in the Securities Holding Statistics Groups database (SHS-G) for individual banks and groups of banks. Additional data sources for granular data are also market data providers, like, for example, securities prices provided by Bloomberg or ratings of securities and issuers from rating agencies.

bank, or the same identifier might be associated to names and addresses spelled out differently; in some cases, identifiers cannot be related to the group structure of the bank. Hence, these information need to be compared and matched. Given that there are different master data sources, it will not be possible to have a perfect match in all cases where the information provided is incompatible. Hence, the quality of integrated master data hampers the correct identification of the bank for which we want to integrate securities data.

3.2. Semantic integration

In a second step, after identifying the bank, we identify the securities that enter the asset and liability sides of the bank's balance sheet. This step is illustrated in **Table 1.B**. A Single Data Dictionary (SDD) supports our search of definition of concepts associated to the securities stored ISIN-by-ISIN in the CSDB for securities issued and in the SHS-G database for the securities held by the bank. Definitions coming from different sources are mapped within the SDD. For example, the 2-digit ISO code defining the country of issuance of the security is mapped to country codes saved in other databases. However, if the definitions of country codes related to securities data are not congruent across different data sources, it is not possible to map all codes. Only with full mapping of the country codes stored in the different sources we can identify all securities issued by the same bank in the same country. In other words, discrepant definitions across databases or mapping errors can affect the quality of the semantic integration.

3.3. Metadata integration

In order to understand the methodology of collection, compilation and dissemination of data, we consult the associated metadata. Metadata values and descriptions are commonly included in data attributes.[6] The application of global standards to structured metadata[7] allows retrieving the information in a straightforward manner. It is less so for data that do not respond to global standards, for which we need to assess, evaluate and compare information about format and methodology. Integrating metadata is therefore necessary to support the reconciliation of data formats and compilation methods. For example, as shown in **Table 1.C** securities prices might present different frequencies and valuation methods (e.g. averages across the periods, end-of period values). Knowing which frequency and methodology of evaluation are applied across different data sources will help understanding how to combine them and apply them to the securities information. Hence, disposing of structured, complete and accurate metadata influences sensibly the quality of the data integration processes.

3.4. Technical integration

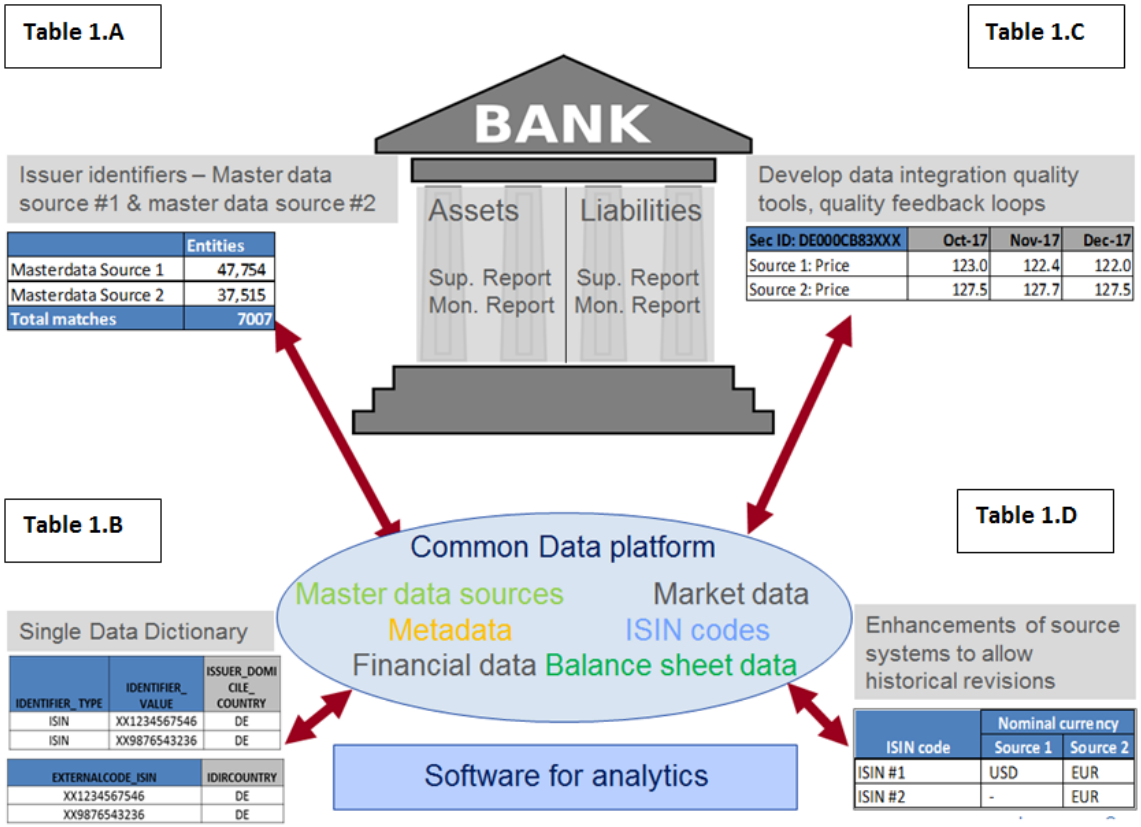
Technical integration is the combination of the IT infrastructure, i.e., hardware and software, to allow for data integration. As shown in the middle of **Figure 1**, a common platform (DISC) provides physical storage to data on-boarded from different sources (internal and external sources for data, master data and metadata). A Data Catalogue and the Single Data Dictionary (SDD) provide the inventory of available datasets and conceptual definitions. The platform should allow users to access queried data and provide feedback on source data quality to data producers, as mentioned in **Table 1.C**.

[6] Metadata could cover several data- related information, including logical, physical and technical information. (DAMA BoK 2nd ed. p.437)

[7] For example, the Standard for Data and Metadata eXchange (SDMX).

Data producers should be able to integrate source data and enhance the quality of the final data product by offering a holistic view of the data architecture. For example, a data lineage should describe origin and changes occurred to the data over time. Furthermore, the technical tools should enable quality assurance processes, for example, by tracking data history and versioning, that is, the evolution of the data along the time and from one database to another due to manipulation, estimations, enrichment and preparation for dissemination, as mentioned in **Table 1.D**.

Figure 1: Data integration for the FAB project on the DISC platform



4. DATA FITNESS FOR INTEGRATION

Each of the activities involved in data integration processes require to have a similar level of maturity to enhance the quality of integrated data. Ideally, all activities should be conducted following a common data strategy and by having centrally designed processes, coordinated policies and data management activities. At this level, an overall increase of data quality is expected[8]. The latter can be ultimately measured trough Data Quality Assurance procedures applied to ECB's official statistics and assessed by using the ECB Statistic Quality Framework (SQF). However, when applying Data Quality Management (DQM) on integrated data, additional dimensions need to be considered, as it is shown in our use case.

In the FAB project, we need to ensure consistency and comparability between granular data and macro data. Therefore, quality checks assess the completeness and consistency of SHS data internally (internal consistency) and vis-à-vis other data sets (external

[8] This is described as level 3, or "defined" level of maturity in DAMA BoK 2nd edition

consistency). For example, quality checks on granular aggregated Securities Holding Statistics (SHS) for a group of banks can be performed against existing aggregated data such as Financial Report data for banks (FINREP). Ideally, matching tables should be implemented centrally and maintained so that the two data sources can be compared on a regular basis. Additionally, ISIN-by-ISIN prices and yield distributions sourced from the CSDB can be compared to data provided by external sources both at the granular observation level and over time.

Validation checks are implemented to verify whether the information on the held securities is correct. For example, the total amount of securities issued and held retrieved from the SHS data need to be lower than the total outstanding amounts shown in the CSDB data. Similar issues relate to the currency of issuance, the sector of issuer, the maturity, the seniority, the rating or other securities related information. Additionally, the adoption of integrity rules allows inquires on the group structure to check if data sources show consistent banks' groups hierarchies. For example, SHS data include holdings of the banking groups and of their individual subsidiaries. The check compares whether the data are hierarchically consistent for each banking group, i.e. whether the sum of sub-institutions equals (or is smaller) than total holdings of the parent and whether the hierarchical structure is consistent with the one of aggregated FINREP reporting.

All these checks, which fit well with the quality assurance dimensions defined for ECB's statistics[9], need to be complemented by additional automated activities for data integration processes, currently performed by data users. These include:

a) **Data customization**: customising and linking data and integrating DQM processes help users receiving the necessary information to perform aggregations. This ensures that the experts' knowledge is passed on. For example, in the FAB project, users might want to re-create indicators provided by commercial data sources (e.g. iBOXX) with prices drawn from the CSDB or from high-frequency sources and compare them to the original data.

b) **Data flagging**: users might need to flag data observations with specific attributes to indicate changes to the status of securities data. This might arise, for example, when securities change their status because of M&A activities of the issuance company. In this case the user might need to add or subtract securities in the aggregation process.

c) **Data derivation**: this activity requires rules to compute derived data from data originated from different sources. The rules need to be defined in comprehensive metadata documentation readily available to data users.

d) **Issue resolution**: users should be able to use feedback loop processes to signal data quality issues to data producers. This would allow data producers to collect feedbacks on their datasets. The adopted issue resolution should be tracked in a dedicated system.

A direct follow-up of these examples has been the formulation of a principle in the ECB data management strategy stating that "data quality and standards must be defined and managed consistently across its lifecycle". This principle has been further developed into

[9] The quality dimensions defined by the Statistics Quality Framework and Quality Assurance Procedures quality for the statistical output include: relevance, accuracy and reliability (including stability), consistency, accessibility and clarity, timeliness (including punctuality).

policy statements that will benefit technical requirements for tools and the processes enhancing data quality assurance and metadata management.

5. INCREASING THE MATURITY LEVEL OF DATA MANAGEMENT ACTIVITIES TO REACH HIGH QUALITY INTEGRATED DATA

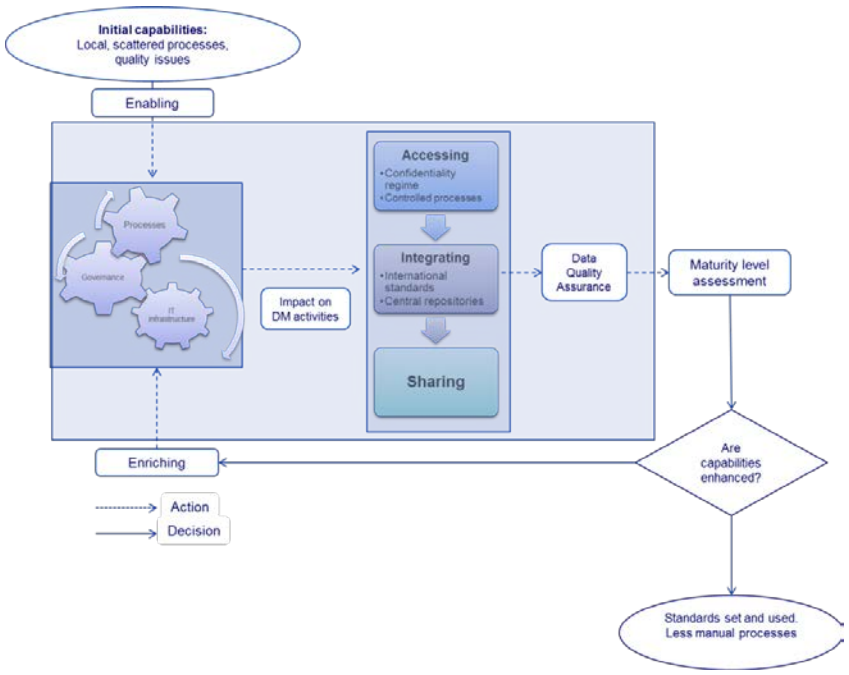
Our use case helps to understand that high quality of integrated data products cannot be obtained without a compatible and defined level of maturity in data management activities that are interrelated with each other. Increasing maturity of data management activities facilitates the gradual abandoning of local manual processes and gradually pushes for the adoption of processes governed by commonly agreed rules and best practices. **Figure 2** helps describing the iteration process for the identification and improvement of the current maturity of data management activities.

The iteration. The FAB project has shown that current granular data integration activities are performed with manual processes raising quality issues. In order to reach a state where processes are more automated and make use of standards, an iterative approach is proposed. As illustrated in **Figure 2** the iteration cycle starts with enabling factors: IT infrastructure, processes and governance. These factors impact data management activities which final aim is to share high quality data products. The latter are not only affected by source data, master data and metadata, but also by the maturity level of the enabling factors. At each iteration data management activities converge into coordinated best practices. All data management components are supposed to have compatible levels of maturity, being interdependent. At each higher level of maturity higher data quality is expected. This is measured through quality assurance processes applied to data, metadata and master data. The iteration ends when the next level of capabilities is reached.

Enabling factors. The first iteration starts with enabling the necessary IT infrastructure as well as initial processes and minimum rules applied to data access, data integration and data sharing. Learning from the experience gained in previous iterations and from the maturity level reached new steps towards higher maturity levels are identified. For example, the FAB project highlighted the need for a common strategy for data integration and for the setting of initial policies for metadata, master data and data quality management. The enhancement should apply to all data management activities, that is, enriching IT tools and processes and reviewing the governance.

Data management activities. Several data management activities are necessary to allow for data integration and data sharing. We represent them with a data process that starts with accessing data, metadata and master data and is followed by their integration and sharing. The FAB project showed initial limitations due to different confidentiality regimes applied to the original data sources. Accessible data are pre-condition to integration activities; therefore, a defined and developed regime for internal data sharing which preserves data confidentiality is necessary. Our use case has also highlighted that integrating data requires the application of standards and the use of central repositories for the efficient management of data, metadata and master data originated by different sources. The ultimate goal of integrating data is to share high quality data products for analytical purposes. Quality assurance procedures are there to measure whether the data product is fit for purpose. The more compatible is the maturity of data management activities, the higher is the quality of the product resulting from integrating data sources. In all stages of access, integration and sharing, new technical requirements reflect more advanced integration processes with the aim to bring IT tools to the state-of-the art.

Figure 2: Increasing maturity level of data management activities



6. CONCLUSIONS

We have presented a use case related to the integration of granular data on banks' securities information to show what is needed to ensure fitness for integration. In other words, integration activities need compatible levels of maturity to ensure high quality of integrated data products. These activities encompass not only quality assurance of source data, but also quality assurance of integrated processes involving: master data and metadata, methodological definitions and technical layers. Each of these areas of data management needs to be enriched in a gradual approach. Policies for sound governance of data quality and metadata management activities enable a change towards more efficient processes in line with the ECB data management strategy of increasing data assets' value. Further coordination in the definition of common rules for best practices is needed in order to develop control on data integration processes that are performed locally and with manual processes. In particular, the ECB data management strategy needs to give priority to the preservation of data confidentiality while promoting data sharing; integrated master data are needed to ensure the successful integration of granular data; finally, the enhancement of state-of the-art IT tools are needed to support efficient data integration processes.

REFERENCES

[1] B. Cœuré, Setting standards for granular data, Opening remarks by Mr Benoît Cœuré, Member of the Executive Board of the European Central Bank, at the Third OFR-ECB-Bank of England workshop on "Setting Global Standards for Granular Data: Sharing the Challenge", Frankfurt am Main, 28 March 2017 available at: https://www.bis.org/review/r170329d.htm

[4] B. Cronin, Thinking about data, Journal of the American Society for Information Science and Technology (2013), 64(3), 435–436.

[5] DAMA International, DAMA-DMBOK: Data Management Body of Knowledge (2nd Edition), (2017), Technics Publications.

[6] European Central Bank, Public commitment on European Statistics by the ESCB, (2013), available at www.ecb.int/stats/html/pcstats.en.html

[7] European Central Bank, ECB Statistics Quality Framework, (2008), available at https://www.ecb.europa.eu/stats/ecb_statistics/governance_and_quality_framework/html/ecb_statistics_quality_framework.en.html

[8] European Central Bank, Quality assurance procedures within the ECB statistical function, (2008), available at https://www.ecb.europa.eu/stats/ecb_statistics/governance_and_quality_framework/html/ecb_statistics_quality_framework.en.html

[9] Eurostat, European Statistics Code of Practices, (2017), available at: http://ec.europa.eu/eurostat/documents/64157/4392716/Revised_CoP_Nov_2017.pdf

[10] M. Lenzerini, Data Integration: A Theoretical Perspective, (2002), available at https://www.researchgate.net/profile/Maurizio_Lenzerini/publication/220266329_Data_Integration_A_Theoretical_Perspective/links/00b4952319d9712541000000/Data-Integration-A-Theoretical-Perspective.pdf?origin=publication_detail