EUROPEAN CENTRAL BANK

EUROSYSTEM

Challenges and Opportunities in Aggregating Business Register Information from Multiple Sources

Ilaria Febbo, Francesca Micocci, Sara Thijs, Alexandru Georgescu, Lorenzo Maria Levati, Nicolò Pegoraro, Thomas Gottron[1]

Abstract

The European Central Bank's Register of Institutions and Affiliates Data (RIAD) is a shared register. Its shared character stems not only from the user side, but also from the data providers side. The information in the register is contributed and maintained by multiple sources, mainly by the national central banks in the European Union. The technical system underlying RIAD allows for any source to contribute to any information on any entity in the register. The deliberate decision to design the system to allow for flexible contributions from multiple sources brought along both challenges and opportunities.

The primary challenge is the task of aggregating information from multiple sources into a single view of truth at any given point in time. This includes the storage and collection of all the information on an entity, the harmonization of the information and the resolution of contradicting information. The main opportunity is a data quality improvement, in terms of better coverage of information, improved timeliness of updates and freshness of the data as well as overall higher accuracy.

The paper at hand presents the data governance setup as well as the technical mechanisms for aggregating multiple sources in RIAD. Furthermore, it reports on practical experience with the system and provides an empirical study and analysis of the information provided by the multiple sources. In particular, the analysis gives insights in the increase of information coverage and a quantitative measure of the benefit of collecting information from a multitude of sources.

1 Introduction and Background

The European Central Bank (ECB) maintains a central and shared Register of Institutions and Affiliates Data (RIAD). RIAD serves multiple clients and users as master data set on entities. Therefore, it is a shared data set on the clients' side, serving multiple purposes such as statistics, market operations, financial stability and

[1] Analytical Credit and Master Data Division, European Central Bank, e-mail: Ilaria.Febbo@ecb.europa.eu, Francesca.micocci1994@gmail.com, Sara.Thijs@ecb.europa.eu, Alexandru.Georgescu@ecb.europa.eu, lorenzolevati@yahoo.it, Nicolo.Pegoraro@ecb.europa.eu, Thomas.Gottron@ecb.europa.eu. The views expressed in this paper are those of the authors and do not necessarily reflect those of the European Central Bank. The authors would like to thank Romana Peronaci and Riccardo Bonci for valuable comments.

banking supervision. All clients benefit from a centrally accessible and maintained register providing high quality data.

However, RIAD is also a shared data set on the side of the data providers, as it collects and aggregates information from multiple sources at national central banks (NCBs) in the European Union (EU) and at the ECB. The technical systems underlying RIAD enable each source to make contributions to every entity stored in the system. The idea behind this approach is that knowledge about newly created entities, changes regarding their information or the relations among entities might become available at different points in time at different sources. Hence, the ultimate aim of bringing all this information together is to ensure high standards of data quality, in particular regarding the completeness, correctness and timeliness of the data. In this way, RIAD currently contains information on nearly 10 million entities and their business historicity aggregated from more than 30 sources.

At the same time, bringing together all this information and serving it to clients in a consistent way also poses challenges. Some sources may contribute only partial information, while others provide old or deviating information. This input needs to be composed into one final and authoritative view on each entity which is served to the clients. To this end, the information is weighted at attribute-by-attribute and source-by-source level to compound it into a final view of what constitutes the current state of truth.

More than one year has passed since the last major release of RIAD's technical system, strengthening in particular the use of multiple sources. In this paper we highlight the setup from a technical and governance point of view, illustrate the aggregation of information from multiple sources and the computation of a single authoritative record. Furthermore, we perform an empiric analysis of the data currently stored in RIAD with particular emphasis on the impact of multiple contributing sources. To this end, we investigated data stored in RIAD under several aspects, indicating the added value provided by the sources. In particular, we could quantify the information obtained from additional sources, the impact of sources on the final, authoritative view and the type of information and entities benefitting most from the technical setup. Furthermore, we have qualitative information on the benefits for business processes and data quality in RIAD.

The rest of the paper is structured as follows. We describe aspects of data governance and technical implementations in Section 2. In Section 3 we describe and present the results of our quantitative analysis on RIAD data, which we discuss and interpret in Section 4. We conclude the paper in Section 5 with a summary and an outlook at future extensions or the RIAD system and further analytical questions.

2 A Multiple Source Approach for Data Collection

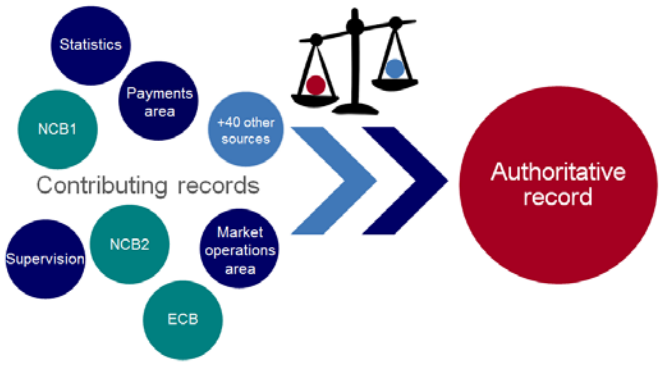
A pivotal functionality of RIAD is the ability to deal with information retrieved from multiple sources. RIAD allows for receiving information from both domestic and non-domestic sources meaning that all RIAD data providers from the European System of Central Banks (ESCB) can contribute reference data to an entity regardless of the

residence of this entity. The RIAD governance framework introduces the concept of competent NCB and contributing NCB. The competent NCB (i.e. the NCB of the respective member state) is ultimately responsible for managing entities which are resident in its member state. For this set of entities, all other NCBs or the ECB are referred to as contributing parties for – from their perspective – non-resident entities. Thus, the competent NCB also needs to manage how information received by contributing NCBs is condensed into one single view.

To enable data provision on both resident and non-resident entities, RIAD offers a wide variety of domestic and non-domestic sources. Competent NCBs can provide data for resident entities under 14 domestic (or national) sources. These sources represent different business areas within the NCB (e.g. statistics, market operations, banking supervision) or country-specific sources. This setup enables NCBs to align internal data sourcing with data provision in RIAD. However, it is up to each NCB to coordinate which sources are used and which information ends up under each domestic source according to the local governance. In addition, RIAD foresees for each European NCB and the ECB one non-domestic source where contributing NCBs can complement information for non-resident entities.

Such a rich data sourcing by multiple data providers requires a prioritisation of information under the different sources in order to generate a unique and reliable authoritative view. While all original information remains in the data set under their respective source, the domestic NCB should decide which information is considered the most correct one and include this as the authoritative view (cf. **Figure 1**). Given the high volume of data in RIAD, it is not feasible to make this decision on an individual case-by-case basis. Rather it is automated using a prioritisation mechanism.

Figure 1 Information from multiple sources is compounded into an authoritative record



This prioritisation mechanism is referred to as ‘compounding’. For each attribute the competent NCB can calibrate which information ends up in the authoritative record according to a specific hierarchy of the different domestic and non-domestic sources. Such an attribute specific hierarchy is called a ‘compounding rule’. The compounding rules to derive the respective authoritative values can be different for each attribute and can be changed over time by the competent NCB. They are triggered every time a data provider makes an update of a value in RIAD. As values in RIAD are

uploaded with a specific validity range, the authoritative records can consist of different source values at different points in time.

3 Systematic Analysis of Multiple Source Contributions

The main purpose of collecting information from multiple sources is to improve data quality. Sources may complement and correct information for each entity on a single attribute level. The authors aim to explore to what extent multiple sources are used for different sets of entities and attributes and how different data sources may impact the final authoritative record. This section describes the results of an empirical analysis performed on a data set extracted from RIAD[2]. The key characteristics of the data set are described in **Table 1**.

Table 1 Characteristics of the analysed data set

Data snapshot	August 2019
Number of entities	9,270,892
Number of analysed attributes	22
Number of sources	36
Number of observed contributions	106,980,270

The analysis is performed along the following set of indicators:

1. Contribution of each source to the authoritative record: how often information coming from a specific source ends up in the authoritative record. This indicator provides information on which source is more prominent in the authoritative record, taking into account the weighing of different sources according to the compounding rules.

2. Contributing source breakdown by attribute type: how many different sources are used for each attribute type. This indicator provides a notion of which attributes benefit more from the multiple source approach.

3. Number of contributing sources per entity and attribute: how many different sources are used for each attribute of a distinct entity. This indicator enables exploring the distribution of the number of different sources per attribute at entity level (considering e.g. minimum, median, average and maximum). It provides insights in which entities are benefiting more from the multiple source approach and to what extent it is actually applied across the entire data set.

4. Ratio of information complementing the main contributing source: for each entity we consider the number of attributes provided by the main contributing source and compare it with the total number of attributes. This ratio is a proxy for the volume of additional information in RIAD because of the multiple source approach.

[2] The data set focuses on attributes that are of relevance for multiple RIAD users. Other attributes in RIAD are often use-case specific (serving only specific users) and therefore a bias may exist in how values are registered and maintained

In addition, we analyse the indicators for the following breakdowns:

- Institutional sector (European System of Accounts classification, ESA 2010)
- Country (EU versus non-EU countries)
- Source (domestic versus non-domestic sources).

In particular, the ESA sectors allow for distinguishing between financial and non-financial entities. **Table 2** outlines the volume of entities in RIAD per sector and shows that non-financial institutions represent 96% of the data set.

Table 2 Breakdown of entities in RIAD by sector

Type of entity	Number of entities
MFI	8,939
Non-financial	8,835,014
Other financial	392,870

As RIAD was initially set up to collect reference data on financial entities, we expect more contributions from multiple sources for entities in the financial sector compared to entities in the non-financial sector. More specifically, we focus in this paper on Monetary Financial Institutions (MFIs) versus non-MFIs due to their relevance for the ECB.

The country breakdown enables the analysis on the use of multiple sources along the RIAD governance. For each EU country, the competent NCB is responsible for maintaining data on resident entities. For other countries, the ECB is competent for managing data on non-EU entities. As non-EU entities are usually registered by European NCBs, the ECB needs to rely on information provided by various non-domestic sources. We expect a more frequent use of non-domestic sources for non-EU entities compared to EU entities.

The source breakdown allows for distinguishing between entities where information comes mostly from domestic sources versus non-domestic sources. We expect that resident entities registered by the competent NCB have fewer contributions from non-domestic sources compared to entities registered cross-border by a contributing NCB.

3.1 Contributions of each source to the authoritative record

The analysis of different source contributions to an entity's authoritative record reveals that 99.2% of the authoritative information is provided by domestic sources, while 0.8% is provided by a non-domestic one.

It is relevant to mention that most NCBs favour domestic sources over non-domestic ones in their compounding rules. Hence, non-domestic sources typically contribute to the authoritative record only if no information is available from domestic sources. Thus, while from a relative perspective the impact of non-domestic contributions may seem low, the situation looks different when looking at the absolute numbers. The

observation means that for 850,000 attributes the competent NCB did not have information. Accordingly, no authoritative record would be available if only information from the competent NCB would have been used.

Focusing on the ESA breakdown shows that for MFIs the impact of the multiple source approach is more substantial: 22% of authoritative information comes from a non-domestic source. This can be explained by the fact that EU MFIs are relevant for multiple data sets for which RIAD hosts entities and, thus, many data providers act on these entities. For example, RIAD hosts the full reference population for the Analytical Credit data set. As one EU MFI might be creditor or debtor to loans in various countries, all respective countries will report or contribute to this MFI in RIAD.

Breaking down the results by country shows that the impact of the multiple source approach is even higher for non-EU entities. While for entities resident in an EU country only 0.8% of authoritative information is provided by non-domestic sources, for non-EU entities the ratio is 85.5%. The explanation for such a difference is that for EU countries the majority of the resident entities are registered directly by the competent NCB under a domestic source. On the other hand, most non-EU entities for which the ECB is competent were registered by a contributing NCB using a non-domestic source. Hence, as long as no new or updated information is provided by the ECB, the information from non-domestic sources prevails.

3.2 Contributing source breakdown by attribute type

Analysing the amount of sources used for each attribute type, we understand that some attributes benefit more from non-domestic contributions than others. Typical examples are the balance sheet related attributes. RIAD records information on annual turnover, total balance sheet and number of employees twice: once excluding and once including foreign branches, if any. Unsurprisingly, the measure which includes foreign branches benefits more from non-domestic contributions: NCBs from countries where headquarters reside might complement the information of the branch (cf. **Figure 2**).

An interesting result comes from the analysis of the attribute types when considering only MFIs. For MFIs the identification properties (e.g. name, street, city), have a significantly higher number of contributing sources compared to other attribute types (cf. **Table 4**). This reflects the fact that such information might be easier to collect.

Figure 2 Contribution of RIAD non-domestic sources to balance sheet related attributes

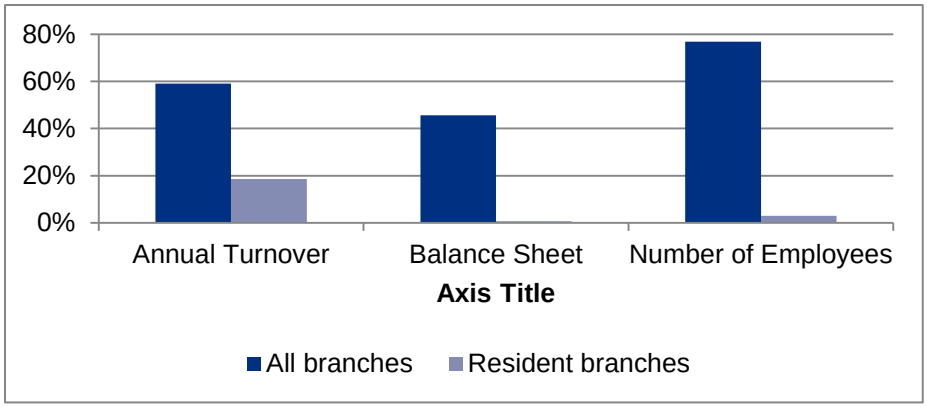


Table 3 Average number of contributing sources by attribute type for MFIs

Attribute name	Average number of contributing sources
Entity name	3.07
Economic activity	2.59
ESA Sector	2.57
Street	2.20
City	2.16
ESA Sector control	2.14

3.3 Number of contributing sources per attribute and entity

In order to understand the positive effect of the multiple sources approach, we calculate an indicator based on the number of contributing sources per attribute at entity level. **Table 3** illustrates that in RIAD we observe attributes of entities receiving information from up to 16 different sources. However, the average number of contributing sources per attribute is 1.02. This means that most of the attributes of the entity are populated by only one source. Looking at the summary statistics we can conclude that the variable number of sources per attribute follows a power law distribution, with most of the observations concentrated in the left part of the distribution and a long right tail (cf. **Figure 3**).

Again, the stratification by ESA sector reveals a significantly different picture for MFI attributes when compared to the other categories. For MFIs, indeed, the number of sources per attribute varies in a range between 1 and 16, with an average number of sources per attribute equal to 1.92 and a median equal to 2. This means that on average MFI attributes are populated by 2 different sources. This can be attributed again to the fact that MFIs are relevant entities for all NCBs providing information to RIAD, which results in multiple NCBs contributing the same type of attribute on an entity.

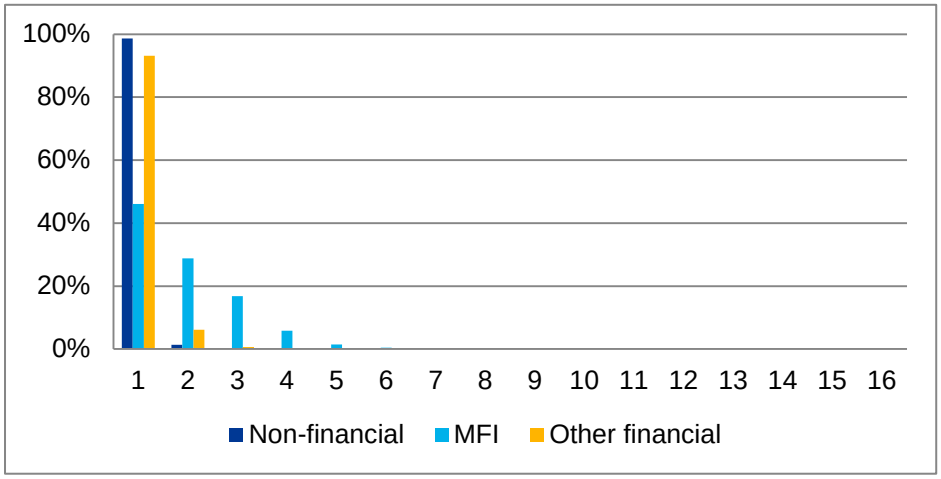
On the other hand, the numbers are quite small for non-financial and other financial institutions. This can be explained by the fact that most of these entities were

uploaded only recently to RIAD[3]. The data stored on these entities consists primarily of the initially uploaded information, typically coming from a single information source.

Table 4 Summary statistics on the number of Sources by entity and attribute, stratifying by entity ESA sectors

	Min	Quartile I	Median	Quartile III	Max	Mean
MFI	1	1	2	3	16	1.93
Other financial	1	1	1	1	7	1.08
Non-financial	1	1	1	1	10	1.02
Total	**1**	**1**	**1**	**1**	**16**	**1.02**

Figure 3 Distribution of number of sources by entity and attribute, stratifying by entity ESA sectors



3.4 Ratio of information complementing the main contributing source

Another indicator for the benefits of the multiple source system is represented by the ratio of information complementing the main contributing source. This indicator is calculated as 1 minus the ratio between the number of attributes coming from the main contributing source and the total number of attributes assigned to an entity as result of the compounding mechanism[4]. The result is a percentage that measures the additional information coming into RIAD based on the multiple source concept. As outlined in **Table 5**, the indicator reveals for non-financial entities, that on average the ratio of information complementing the main contribution source is 4.82%. This result, together with the average number of sources contributing to the authoritative record per entity, is in line with the observation that for non-financial entities the attributes are generally populated by a single main source. In line with the previous

[3] With the go-live of AnaCredit in 2018 the volume of entities registered in RIAD has multiplied nearly by a factor of 20.

[4] Please note, that for this analysis we consider all sources, independent of them being domestic or non-domestic

paragraph's results, for MFIs the ratio of information complementing the main contributing source is much higher: 23.24%.

Table 5 Impact of the multiple sources concept on information

	Non-financial	MFI	Other Financial
Ratio of information complementing the main contributing source	4.82%	23.24%	7.93%
Average number of sources contributing to the authoritative record per entity	1.16	2.90	1.43

4 Discussion

The quantitative analysis presented in the previous section gives an overview of the current state of how the multiple source approach in RIAD is used by the NCBs and the ECB. The indicators show that the source concept is actively used and that the data coverage benefits from the approach to a significant extent. Moreover, there are several observations which are worth a closer look.

We observed that entities in the financial sector benefit much more from multiple sources than non-financial institutions. Firstly, financial institutions have been maintained in RIAD for a much longer time and are of higher relevance for the ECB's monetary and supervisory tasks. For example, different domestic sources may provide information relevant for statistical and supervisory purposes. Secondly, as EU MFIs are relevant for multiple (instrument) data sets for which RIAD hosts entities and are involved in many cross-border relationships, also several non-domestic data providers act on these entities. For example information on cross-border relationships or information collected at the level of a non-domestic head of a branch is often registered under a non-domestic source.

For the bulk of recently registered non-financial entities the situation looks quite different. Only a relatively small number of them currently receive information from non-domestic sources. However, we observe the use of non-domestic sources to be growing and, for at least some of this information, we would expect also non-financial entities to benefit more from non-domestic sources over time. While in the non-financial sector we will probably not reach the levels observed on MFIs, this is definitely a development to monitor closely.

For entities not resident in the EU, instead, we observed an extremely high contribution of non-domestic sources. This information at large stems from the fact that such entities are reported by NCBs. Subsequently, the ECB takes over competence for these entities and step-by-step extends and updates the information. Hence, here we would also expect a normalisation of the indicator values to a level where also domestic information sources provide an important share of information.

In general, all observations indicate a strong potential of the multiple source approach. Already now, much information is provided from different sources and more than 5% of the authoritative information in RIAD is only available due to additional sources. The MFIs serve as a best case scenario where the benefits are exemplary for what can be achieved.

The results also motivate a careful revision of the compounding rule settings. On the basis of a future qualitative analysis of the attribute values coming from different sources, we might understand better which sources provide information timely, more complete or more correct. This could support the adaptation of source weights in the compounding rules to give a more appropriate importance to sources at attribute-by-attribute level. We already have anecdotal evidence, that some NCBs are making use of information coming from non-domestic sources by incorporating them in their own national systems. There is evidence for such behaviour in the data, as there are cases where attribute values reported under non-domestic sources are picked up and reported also under domestic sources with a little time delay. This can partly be attributed to the fact that according to the RIAD Governance NCBs are responsible for managing entities residing in their member state and thus, have to correct and complement reference data on resident entities registered by contributing NCBs. A more detailed analysis may quantify this effect and provide further evidence.

Overall, we see many benefits in the multiple source approach. The presented analysis provided interesting insights and will keep a close look at the developments in the future and extend the analysis to investigate further.

5 Summary and Conclusion

In this paper we investigated the impact of the multiple source approach in the ECB's register RIAD. We described the technical setup of collecting information on entities from multiple sources, covering different domestic and non-domestic sources and the compounding rules which are used to infer an authoritative record, constituting the definitive view on the information on each entity. In a data-driven analysis we investigated a snapshot of data in RIAD from August 2019, consisting of information on nearly 10 million entities, more than 100 million attribute values and more than 30 information sources. We found that approximately 5% of the information in RIAD is available because of contributions of additional data sources. This is particularly valid for entities in the financial domain where RIAD has a long tradition of registering information, as well as for entities involved in cross-border business relationships or entities resident in non-EU countries. Furthermore, we observed how some national central banks benefit in their business process, actively incorporating information from non-domestic sources in their local registers.

Future work and a follow up action are to analyse in more depth the quality of the contributions. Quantitative results may hint at potential gains in data quality by considering non-domestic data sources. However, to review if additional data also leads to an increase in quality, we need to analyse the content of the contributed data. This might form the basis for careful adjustments in the compounding rules. Furthermore, a currently considered extension for RIAD is the inclusion of external data sources in the input layer. These data sources could be used to incorporate and align information from other registers at European institutions, non-EU business registers or other data providers. Once introduced, a similar exercise as the one presented in this paper could demonstrate the benefit of each data source.

References

1. European Central Bank (2018) ‘Guideline (EU) 2018/876 of the European Central Bank of 1 June 2018 on the Register of Institutions and Affiliates Data (ECB/2018/16)’. Official Journal of the European Union L 154, 18.06.2018, pp.3-21.

2. Kropp, M., Thijs, S., Neudorfer, P. and Corvoisier, S. (2017) ‘Connecting the ESCB Business Register with Other Supranational Sources – the Example of Linking up to the GLEIF Platform’. Meeting of the Group of Experts on Business Registers, European Central Bank, Paris, France, 27–29 September 2017.

3. Neudorfer, P. (2010) ‘The Role of Administrative Sources in the ECB’s ‘Register of Institutions and Assets Database’ (RIAD)’. 22nd Meeting of the Wiesbaden Group on Business Registers, European Central Bank, Tallinn, Estonia, 27–30 September 2010.

4. Neudorfer, P. (2013) ‘Managing the Quality of the ECB’s Enhanced ‘Register of Institutions and Affiliates Database‘ (RIAD)’. Meeting of the Group of Experts on Business Registers, European Central Bank, Geneva, Switzerland, 2–4 September 2013.

5. Neudorfer, P. (2014) ‘Integrated Data Collection and Reporting in the European System of Central Banks (ESCB) and the Single Supervisory Mechanism (SSM) – The Role of the “Register of Institutions and Affiliates Database” (RIAD)’. 24th Meeting of the Wiesbaden Group on Business Registers, European Central Bank, Vienna, Austria, 15–18 September 2014.

6. Neudorfer, P. (2016) ‘Multiple Usage of the ESCB 'Register of Institutions and Affiliates Database' (RIAD) – Features and Challenges’. 25th Meeting of the Wiesbaden Group on Business Registers, European Central Bank, Tokyo, Japan, 8–11 November 2016.

7. Perrella, A. and Catz, J. (2018) ‘Interconnecting Multiple Granular Datasets to Evaluate Credit Risks – The ESCB Experience’. 17th International Conference on Credit Risk Evaluation Designed for Institutional Targeting in Finance, European Central Bank, Venice, Italy, 27-28 September 2018.

8. Thijs, S. and Corvoisier, S. (2017) ‘Integrating Reference Data for Monetary Policy and Supervisory Purposes - The European System of Central Banks (ESCB) Experience’. IFC Bulletin 43(5), pp. 1-12, Basel: Bank for International Settlements.