11th Biennial IFC Conference on "Post-pandemic landscape for central bank statistics"

BIS Basel, 25-26 August 2022

Data quality management of entity group data: relevance and tools to address current challenges[1]

Bruno Carreiras,
European Central Bank

[1] This presentation was prepared for the conference. The views expressed are those of the author and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the event.

Data quality management of entity group data: relevance and tools to address current challenges

Bruno Carreiras[1]

The COVID-19 crisis brought renewed challenges for statisticians in Central Banks, especially those dealing with granular datasets. This paper addresses some of those challenges in the context of entity group data by proposing a metric to identify "relevant" changes in group compositions (joiners and leavers) and rank them by importance. We show how such an indicator can provide a useful tool for an effective prioritisation of data quality management activities within the statistical community. As a complement, we also show how the same indicator can be used as an evaluation technique for corporate changes.

Keywords: granular data, data management, data quality, groups, master-data, statistics, network data, central banking.

Contents

[1] European Central Bank, DG-Statistics, Analytical Credit and Master Data Division, Research Analyst. I would like to thank my colleagues from the RIAD team in DG-Statistics for their input and guidance. The views expressed in this paper are those of the author and do not necessarily reflect the views of the ECB.

Introduction

Data quality is one of the most crucial topics in the world of increasing data volumes in central banking[2]. As datasets become more granular and data is collected more frequently, data quality issues increase and can be more difficult to assess. This combined with the need for accurate statistical indicators in real time, amplified with the recent COVID 19 crises (De Beer, Tissot 2020), creates an **environment where data must be managed efficiently.**

Some argue that innovation can be a tool to respond to these challenges[3]. The usage of machine learning and other innovative techniques to manage data quality has already proven useful to some of the ECB's granular datasets (see Romano et al, 2021). This work intends to be an addition to that innovative wave by proposing a new way to efficiently prioritise the data quality management of the entity group structures stored in the RIAD[4] the ECB master data platform on entities.

RIAD groups are 'network trees' (company structures) which are generated each day based on relationship data between entities. These groups are monitored every day for any changes in their structure (movements) and any quality issues are followed up. With the steady increase in the number of relationships and consequently in the number of groups, more movements need to be assessed by data quality managers. Under this context, it is **important to know which group movements are the most relevant so to follow-up in an efficient manner**.

This paper tries to respond to that need by proposing a metric, the **Significance Multiplier (SM)**, that **ranks movements inside group structures according to their impact on supported business processes**. This metric replies on an input-based model, where users can decide "what is important for them", depending on their specific loss function. As a by-product, the same tool allows **evaluating group changes in corporate structures motivated by real business cases** (e.g. mergers or acquisitions). By being a complementary calculation to the group structure and an addition to the dataset, the metric can also be refreshed on a daily basis. Thus, it can also be used to identify *a priori* the impact of certain movements and support decision makers.

The paper is structured as follows: in Section 1 the RIAD dataset and the RIAD group structures are described. In Section 2, the data quality checks that are applied to RIAD relationship data are laid out. Section 3 introduces a baseline concept for group importance/relevance and related data quality prioritisation. Moreover, that same baseline concept is applied into the context of the group members and movements. Still in Section 3, the baseline scenario is extended to the input-based approach that aims to internalise specific business needs in the significance multiplier.

[2] "Don't take it for granted: the value of high-quality data and statistics for the ECB's policymaking"- Schnabel (2020)

[3] The potential role of innovation is mentioned in Rosolia et al (2021) or De Beer, Tissot (2021)

[4] Register of Institutions and Affiliates Data, see Background section for more information.

In Section 4, two applications are laid out, related to data quality prioritisation (via priority ranking) and to the evaluation of movements under a specific business case. Section 5 draws the main conclusions and puts forward possible future extensions.

1. Background

The Register of Institutions and Affiliates Data (RIAD)[5] is the ESCB shared master dataset on entities serving several central banking and Single Supervisory Mechanism (SSM) (even daily) business processes and statistical data collections. RIAD is operated in the ECB's Directorate General Statistics and provides master data (e.g. name, address, legal form, institutional sector) on various types of Organisational Units (legal entities and their affiliates, such as branches and subsidiaries). An important feature of RIAD is that it provides the relationships (bilaterally) linking between the entities.

RIAD data is collected from National Central Banks (NCBs) and other National Competent Authorities (NCAs)[6] compliant with several legal acts[7], and provided by the latter based on various data sources available at national and international level.

The RIAD system allows data providers to create, change and delete data (including relationships) on any Organisational Unit (or 'entity', hereinafter) on a daily basis. This flow provides the flexibility needed for a central system to gather information coming from around Europe and swiftly replicate business changes, hence providing always up-to-date information needed for the performance, *inter alia*, of key central banking activities (e.g. risk management and collateral management). **Based on (entity-to-entity) relationship data provided by all relevant sources, RIAD calculates (different types of) group structures every day**.

RIAD groups are based on navigation algorithms that start from the valid relationships collected in the system and generate a 'network tree structure'. Figure 1 shows how a typical RIAD group network could be represented graphically.

These algorithms were designed to support various stakeholders with respective business processes: as a result, there are currently four different group calculations in RIAD[8], they are used to perform tasks such as assessing the links between institutions

[5] As defined in Guideline (EU) 2018/876 of the European Central Bank on the Register of Institutions and Affiliates Data (RIAD)

[6] National competent authorities in the context of SSM data collection and as defined in: Guideline (EU) 2020/497 of the European Central Bank on the recording of certain data by national competent authorities in the Register of Institutions and Affiliates Data

[7] RIAD data provision is based on the following legal acts: Guideline (EU) 2018/876 of the European Central Bank on the Register of Institutions and Affiliates Data (RIAD); Guideline (EU) 2019/1335 of the European Central Bank amending Guideline (EU) 2018/876 on the Register of Institutions and Affiliates Data (RIAD); Guideline (EU) 2020/497 of the European Central Bank on the recording of certain data by national competent authorities in the Register of Institutions and Affiliates Data

[8] For this paper only the algorithm based on ownership relationships is analysed, any reference to number of groups only relates to that algorithm. In short, those groups are calculated in the following way: if entity A has an ownership relationship towards two entities (B and C), RIAD would calculate a group structure where A is the head and B and C are its children. If B would own another entity D, then the group structure would comprise four entities (adding one to the bottom). Additionally, one-to-one relationships are also considered a group (with two members).

pledging collateral for monetary policy operations, monitoring supervised banking groups, or calculating consolidated (loans/securities) exposures. As a result, any data quality issue affecting RIAD groups data (and, before that, the underlying relationships) may negatively affect such important tasks.

Currently there are ca 130,000 groups in RIAD, based on a universe of ca 300,000 relationships. As explained above, any changes on the universe of relationships may create movements (genuine or not) **in the group structures**. Assessing the changes in such groups, and ensuring data quality, is a clear challenge given limited resources and user-specific priorities.

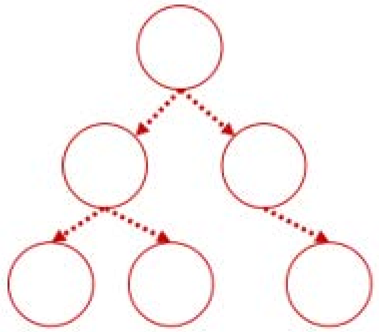


Figure 1 - Graphical representation of a RIAD group structure.

2. Quality assurance of group data in RIAD

Relationships in RIAD (the source data for group calculation) are subject to automated quality checks embedded in the system and serving as a check for the consistency, completeness, and accuracy of the data. Relationships which do not pass such controls are not considered in the calculation. On the other hand, valid relationships enter the daily calculation and generate the group population.

It often occurs that the daily group calculation is affected by the dynamic nature of the RIAD collection environment. At any point in time, an entity may leave or join a group and create a change in its structure. These are the so-called **group movements**. A movement is triggered by a change in the underlying data that can be either genuine or erroneous. This brings the need for an extra layer of quality checks on groups**: post-calculation monitoring**.

Let's take as an example the case of a new relationship being uploaded into RIAD on a certain day: if the new relationship is added to the top of a group signalling a new head, it can be assessed either of two ways:

1. This change is legitimate and in fact the group has a new head (e.g. a new holding at the top of the group).
2. This change is not legitimate and reflects a mistake in the upload of a relationship (e.g. wrong parent/child direction).

Case b) is a clear inconsistency, while case a) is a genuine change. In the end, experience proved that there would always need to be a subjective "eye" that would define legitimacy in both cases.

In a large and growing group population, if a data provider/data quality manager would need to confirm the legitimacy of every movement, there would be a significant operational effort related to group monitoring. To mitigate this effort, it is important to define a **prioritisation measure that can point a data quality manager to which cases should be assessed first**.

3. Modelling relevance and prioritising group movements

3.1 Significance Multiplier to measure relevance of groups and group movements

The number of groups and the complexity of the underlying data are challenging to manage. To efficiently deal with the quality management of such an amount of information there is a need to **classify all potential issues so to allow DQ managers to focus on the most relevant changes/movements**. In other words, there is a need to define a concept of "significance" or "relevance" in the context of RIAD group structures.

The starting premise is that the baseline criteria should be the **size of the group (S) affected**, i.e. number of group members. If $S_{G1} > S_{G2}$, then G1 is more relevant than G2 and data quality should be prioritised for G1. By assessing the size of the group (S_G), one can rank all existing groups by importance and define a baseline prioritisation metric.

However, the size of group networks as a first level of prioritisation can be reductive. The number of members in a group does not necessarily represent the best measure of how relevant a movement inside the group is.

If a group member moves, the group network is affected at every level that is dependent on that group member, i.e. a group member always leaves with all its children (if any). Based on this, one can state that a movement is as *significant* as the group members that it involves. Thus, the **importance of a group member/movement can then be defined as the size of the movement in its group resulting from a single event affecting its own presence**, or ***significance multiplier for RIAD group structures (SM)***[9]. In simpler terms, the SM is the size of the sub-group headed by a certain group member.

Under the assumptions outlined above, looking at figure 2, the exit at T+1 of group member X is less important as it impacts only its own presence (SM=1). While in figure 3, the movement of group member X (as leaver), affects the presence of three entities, signalling that this is a more relevant group member (SM=3).

[9] This metric is an application of a weighted centrality measure for network data to the RIAD data model and its algorithms for a concrete business purpose.

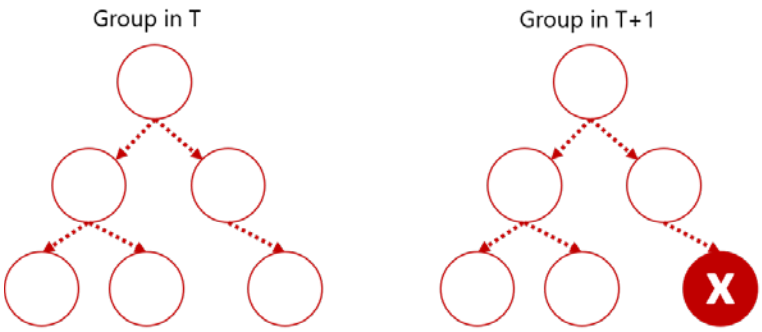
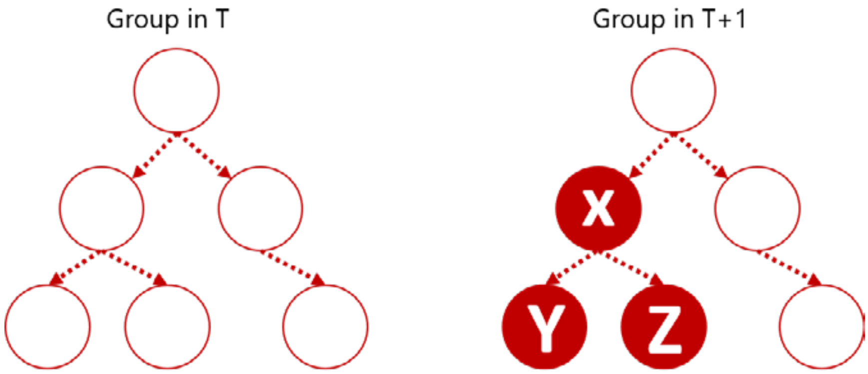


Figure 2 - Movement in a group structures, in filled red the one leaver – SM = 1



.

Figure 3 - Movement in group structures, in filled red the leavers. X affected by a single event, Y and Z affected by association - SM = 3

The SM provides a good measure for *importance* of a member inside the group population, i.e. you can compare entities inside the same group and across groups. If the SM of i in group G is higher than the SM of j in group G, then i is more important than j, by the same logic any movement that *starts* in i is more relevant than any movement that *starts* in j. Under the baseline scenario that the amount of entities affected is what matters, one can also compare the SM across groups. This suffers from the **limitation that groups with more members will necessarily have more significance** (this also applies to its movements). This raises the challenge of finding a metric that is more accurate in defining relevance in the whole population.

3.2 Prioritising and evaluating movements under a dynamic business environment

The usage of RIAD group structures has increased significantly in recent years. Many stakeholders are interested in performing their analysis with RIAD data; besides that, several key central bank functions (e.g. risk management, collateral management) rely on RIAD data for their (daily!) operational processes. All RIAD clients have different expectations and needs and hence different criteria to identify "relevant" groups. As an example, in the context of AnaCredit[10] such relevance may be measured with a credit-related property (e.g. outstanding nominal amount of the loan), while for the

[10] As defined in: Regulation (EU) 2016/867 if the European Central Bank on the collection of granular credit and credit risk data

climate change purposes[11] this could be based on the value of variables such as the carbon intensity or the main economic activities (as a proxy for that). The nature of these differences raises another question: how can one homogeneously apply measures of importance that are suitable to all clients' expectations and business needs while at the same time coping with limited resources in a shared master dataset?

As defined in the previous section, the importance of a group at its very baseline is measured by the number of its group members. In principle, this would provide a very good proxy of relevance for a group. However, it would suffer from the immediate limitation that relevant groups with a small number of members would be deemed as less important. To tackle this, an **extension based on a criterion applied to specific business cases (α)** is proposed. The business specific criteria would be a pre-defined (input based) variable selected by the stakeholder. This criterion should allow to **accurately quantify the importance of a group for a certain business process**.

At full group network level, the **definition of size (S) becomes dynamic**, i.e. size should be measured differently than a pure count of group members but rather with a relevant quantifier (αS_G):

$$\alpha S_G = \sum_{i \in G} \alpha_i$$

Where G is the group structure, i is a group member belonging to G and α is a business specific criteria for a group member. In short, this would mean that all the business specific criteria for every group member would be summed up.

Under these new conditions, it is now possible that a smaller group (e.g. with 20 members) has more importance than a larger group (e.g. with 200 members), if the sum of its α is larger.

Naturally, the same logic can be applied to group members. While in section 3 the SM is dependent on the number of members affected by an event on one given member, under this extension it is dependent on the sum of α affected by an event on one given member.

$$\alpha SM_i = \sum_{i \in G_k} \alpha_i$$

Where G is the group structure, G_k is the sub-group of G impacted by the single event on k. In practice, a group member would be valued as the sum of the business criterion of all the entities it has below it (the "sub-group"). In figure 3, the αSM of X would be the sum of the α of X, Y and Z.

The business dedicated method allows for a clear definition of importance of a member inside the whole group universe by defining a **criteria (α) that provides a universal measure of comparison (cross-group)**. As a result, you can accurately compare two movements (or group members). This is the advantage of the αSM in comparison to the original SM.

All in all, the **αSM for DQM defines relevance of movements by positioning (by importance) each group member in the full universe of group structures**.

[11] New unit in the ECB announced in 2021: https://www.ecb.europa.eu/press/pr/date/2021/html/ecb.pr210125_1~3fc4ebb4c6.en.html

Thus, we are prioritising data quality actions by defining a ranking of importance for every group member. If a group member is touched and triggers a movement, this movement will be ranked by importance, as the αSM incorporates the impact.

The fact that this measure is applied equally to every group member provides a rather important by-product. Besides providing this DQM efficiency gain, it also has a use for different business purposes. **The value that this metric allocates to every group member can, for example, be used to evaluate corporate changes, i.e. mergers, splits, acquisitions, in a wide range of topics** (the α can be modified to fit different business cases). For example, if a branch of a certain group would be acquired by another group, the αSM would provide the value that would move across groups. Thus, the business specific SM metric provides both a priority and an evaluation technique. In the next section it will be laid out how this metric can work in practice.

4. Implementation: Significance Multiplier in practice

4.1 Data preparation and construction of the dataset

To show an application of the model described in this paper, the RIAD dataset was used. For the two use cases below, data from two group snapshots was used, end of January and end of May 2022[12].

Group data in RIAD is modelled as a network tree data, i.e. a dataset that replicates the tree graph and the nodes and edges inside it. The nodes are represented by the RIAD ID (identifier at database level). To complement these data, some additional reference data (RIAD Code, Name, ESA Sector) and balance sheet data was added.

With this dataset prepared, the main dataset[13] which includes the application of the SM model is constructed. This adds three main variables: size of the group, Significance Multiplier (SM) of each group member (and head), business specific SM of each group member (and head).

4.2 Prioritisation of data quality follow-up on movements

The tool used for this example flags all the daily movements in banking group structures (similarly to what is depicted in figure 2). The movements are then given a ranking of importance based on the business specific SM (αSM). The criterion (α) used is the individual (domestic) balance sheet[14] of banking group members. The use of this ranking is the focus of this example.

[12] Note that these data are available in the ECB for those who have sufficient confidential clearing. Meaning that one with the same accesses can replicate the dataset.

[13] Dataset description can be found in the Annex

[14] Collected under Guideline (EU) 2018/876 of the European Central Bank on the Register of Institutions and Affiliates Data (RIAD) – Annex I

In the SM model, the baseline theory is the starting point. To prove the usefulness of the business specific criteria, the measures of the model with the baseline values (pure counts) and the business specific values (sum of α) are compared.

The outcome will be twofold:

- A **ranking at group level**: all group heads will be ranked by importance ($S_G - baseline$, $\alpha S_G - business\ specific$)
- A **ranking at group member level**: all group members (excluding heads) will be ranked by importance ($SM - baseline, \alpha SM - business\ specific$)

The following results apply to the banking group sector following the baseline criteria:

Ranking of RIAD banking groups by Size

Comparison of RIAD group structures on count of group members (baseline) Table 1

Group Identifier	Group Name	Size	Ranking
G1	Banking group 1	841	1
G2	Banking group 2	838	2
G3	Banking group 3	781	3
G4	Banking group 4	765	4
G5	Banking group 5	696	5
G6	Banking group 6	668	6
G7	Banking group 7	363	7
G8	Banking group 8	348	8
G9	Banking group 9	348	9
G10	Banking group 10	298	10
G11	Banking group 11	252	11
G12	Banking group 12	246	12
G13	Banking group 13	221	13
G14	Banking group 14	211	14
G15	Banking group 15	186	15

[1] The above data is anonymised due to confidentiality reasons.

Sources: RIAD group data complemented with author's own calculations

Ranking banking group members by Significance Multiplier

Comparison of RIAD group members (excluding heads of groups) based on significance multiplier (SM) Table 2

Entity Identifier	Entity Name	SM	Ranking
M1	Member 1	559	1
M2	Member 2	556	2
M3	Member 3	425	3
M4	Member 4	371	4
M5	Member 5	247	5
M6	Member 6	240	6
M7	Member 7	235	7

M8	Member 8	225	8
M9	Member 9	220	9
M10	Member 10	214	10
M11	Member 11	194	11
M12	Member 12	180	12
M13	Member 13	176	13
M14	Member 14	166	14
M15	Member 15	165	15

[1] The above data is anonymised due to confidentiality reasons.

Sources: RIAD group data complemented with author's own calculations

Thus, according to pure size (Table 1), *Banking group 1* would be considered the most important banking group as it is the one with the highest number of members (841). *Member 1* would be considered the top group member with 559 entities as its direct or indirect children.

To make a reasonable cross-group comparison, one should investigate the results applied to the banking groups under the business specific criterion:

Ranking banking groups by business specific Size

Comparison of RIAD group structures on its business specific size ($\alpha S_G - balance\ sheet$)

Table 3

Group Identifier	Group Name	αSize (Bln €)	Size	New Ranking	Old Ranking
G5	Banking group 5	> 2,000	696	1	5
G7	Banking group 7	> 2,000	363	2	7
G8	Banking group 8	> 2,000	348	3	8
G230	Banking group 230	> 1,000	137	4	-
G15	Banking group 15	> 1,000	186	5	15
G9	Banking group 9	> 500	348	6	9
G231	Banking group 231	> 500	97	7	-
G1	Banking group 1	> 500	841	8	1
G2	Banking group 2	> 500	839	9	2
G10	Banking group 10	> 100	298	10	10
G232	Banking group 232	> 100	114	11	-
G233	Banking group 233	> 100	50	12	-
G234	Banking group 234	> 100	5	13	-
G224	Banking group 224	> 100	246	14	-
G12	Banking group 12	> 100	45	15	-

[1] The above data is anonymised due to confidentiality reasons.

Sources: RIAD group data complemented with author's own calculations

Ranking banking group members by business specific Significance Multiplier

Comparison of RIAD group members (excluding heads of groups) on its business specific significance multiplier (αSM_i (balance sheet))

Table 4

Entity Identifier	Entity Name	αSM (Bln €)	SM	New Ranking	Old Ranking
M320	Member 320	>2,000	21	1	-
M11	Member 11	>2,000	194	2	11
M322	Member 321	>2,000	2	3	-
M323	Member 323	>2,000	1	4	-
M324	Member 324	>1,000	25	5	-
M325	Member 325	>1,000	5	6	-
M3	Member 3	> 100	425	7	3
M8	Member 8	> 100	225	8	8
M326	Member 326	> 100	163	9	-
M327	Member 327	> 100	31	10	-
M10	Member 10	> 100	214	11	10
M328	Member 328	> 100	88	12	-
M329	Member 329	> 100	68	13	-
M330	Member 330	> 100	104	14	-
M331	Member 331	> 100	16	15	-

[1] The above data is anonymised due to confidentiality reasons.

Sources: RIAD group data complemented with author's own calculations

Notice that with the change in the criteria, even though the pure size is a good proxy for importance (as referred in previous sections), the new criteria (Table 3) provides a different view on relevancy. A good example is the inclusion of *Banking group 231* which has only 97 members but the sum of the balance sheet of its members positions it on the 7th place of the group priority ranking.

On group member level, Table 4 clearly shows the change from the baseline situation. The group members have their importance ranked in a significantly different way via the business criterion. A group movement on any of these "arms" of the group would result in a sharp decrease of the consolidated value of the network. Note that some of the members have a very low number of children below them but still an high importance. This would be the exact impact that could be missed under the baseline scenario or without a priority ranking.

In conclusion, the application of the business specific significance multiplier allows for a better approximation of the importance of movements under a specific business case. Its usage as a data quality criterion allows to identify which movements are most relevant. This is **very helpful for regular data quality procedures with the aim of following-up on possible data quality issues** as it prioritises and protects important group members.

4.3 Valuing changes in corporate structures

With the new international landscape, relationship data on banking groups was changed in RIAD to reflect various events. Namely, for the case of *Banking Group X*, there was a change in structure resulting from the actions taken during the first quarter of 2022.

Applying the SM model to this group, the analysis can be broken down by group member and give them their relative importance inside the group (base on a pre-defined criterion). This will be helpful to understand which of the movements were the most impactful.

By comparing the same group structures in different snapshots (end of January against end of May) one can find the entities that moved away from the group (marked in red in the table 5). To measure these changes, the business specific significance multiplier (αSM) at the time before the movements (end-Jan) was collected. As in the previous example, the domestic balance sheet was used; however, any other quantitative metric could be used. The results are described in Table 5.

Naturally, the movements are always reflected in the αSM of the head of the group as they represent an accumulated loss. The advantage of using the αSM on all the group members instead of looking just into the head is that you can evaluate how relevant were each of the "movers" in the total impact in the group. In this example, the impact is only on the *X5* "sub-group" and is then reflected in two entities: the *X12* and *X10* branches. Objectively, you could infer that the αSM would rank the *X12* branch as the most relevant and the *X10* as the least relevant. Of course, looking into the data subjectively, and as the values are so close together, one could argue that they are virtually of the same relevance. Another important point is that one could, in January, know the full amount of balance sheet value that would leave the group by summing up the αSM of the leaving entities. The big advantage is that any dependent entity (children of these branches) would already be accounted in the measure.

This example shows a movement in a small group but picturing a larger movement in a larger group and spread out across the structure, the αSM could provide the same level of insight. Additionally, if such a metric would be calculated daily, one could have an immediate response to the question: What would be the impact if a certain entity leaves its group? In the example above, what is the impact in group structure X if the *X12* branch leaves the group? With the measure explained in this paper, **such an answer could be given as a data point in a dataset, addressing needs from decision makers.**

How did the Banking Group X structure change? (Table 5)

Evaluating the changes in Banking Group X structure using the business specific significance multiplier (αSM, domestic balance sheet)

Entity Identifier	Group Name	Present in May	αSM (Mln €) (Jan)	αSM (Mln €) (May)	Total loss (αSM, Mln €)
Member X1	X1	Yes	> 300,000	> 300,000	**> - 2,000**
Member X2	X2	Yes	-	-	-
Member X3	X3	Yes	-	-	-
Member X4	X4	Yes	-	-	-
Member X5	X5	Yes	> 5,000	>2,000	**> - 2,000**
Member X6	X6	Yes	> 500	> 500	-
Member X7	X7	Yes	-	-	-
Member X8	X8	Yes	> 0.1	> 0.1	-
Member X9	**X9**	**No**	**-**	**-**	**-**
Member X10	**X10**	**No**	**> 1,000**	**-**	**> -1,000**
Member X11	X11	Yes	-	-	-
Member X12	**X12**	**No**	**> 1,000**	**-**	**> -1,000**
Member X121	**X121**	**No**	**-**	**-**	**-**
Member X13	X13	Yes	-	-	-
Member X14	X14	Yes	-	-	-
Member X15	X15	Yes	-	-	-
Member X151	X151	Yes	-	-	-
Member X152	X152	Yes	-	-	-

[1] The above data is anonymised due to confidentiality reasons.

Sources: RIAD group data complemented with author's own calculations

5. Conclusions

In this paper we propose an approach to assess the relevance of groups and their group members in a large master-dataset on entities. The method is based on the group data model of the ESCB's Register of Institutions and Affiliates Data (RIAD), maintained in the ECB's Directorate General Statistics in close cooperation with EU NCBs and NCAs, and allows efficiency gains in the data quality management of such information. The proposed metric:

- Provides a **prioritisation measure** for data quality management on group structures, allowing data quality managers to focus on the most relevant cases from the point of view of the impact on data users;
- Provides a **business-driven measure to evaluate** important movements inside group structures, allowing experts to know the impact of certain decisions beforehand.

This approach is currently used as a prioritisation tool for data quality management on RIAD group data and has proved valuable in correctly identifying the right priorities in data quality work, which is essential in a context of scarce resources. Besides the two use cases detailed in Section 4, the tool was tested with further business criteria and proved successful in the calculation of the significance multiplier of the group population. Also, it showed that changing the criteria has an impact on how the relevance is distributed. For future studies, one could consider showing a comparison between the results under different business criteria.

The metric also proved useful to value subsidiaries inside groups and show their relative importance. An introduction of such a measure, as a complement to a dataset, could result in efficiency gains for experts across the ECB (and other institutions) when evaluating the impact of real business changes. Another possible extension could be to use this metric to evaluate the impact of corporate actions over time or in the context of the climate change work (e.g. using the carbon footprint as a business specific criteria).

Finally, one could reflect on the possibility of applying this approach to other network data, i.e. group data collected for other statistical datasets. As this tool is based on a network dataset, any other data that has the same structure could essentially use the same technique.

Annex

Dataset variables and description

Table 6

Variable	Description	Original dataset
Group ID	Unique identifier of groups in the RIAD system	RIAD group dataset
Group member RIAD ID	Unique identifier of the entities that belong to the groups	RIAD group dataset
Parent RIAD ID	Unique identifier of the entities that are parents in group structures	RIAD group dataset
Group member RIAD Code	RIAD Code of the entities that belong to the groups	RIAD reference data dataset
Group member Name	Name of the entities that belong to the groups and have a parent	RIAD reference data dataset
Group member ESA Sector	ESA Sector of the entities that belong to the groups	RIAD reference data dataset
Group member Domestic Balance Sheet	Domestic balance sheet of the entities that belong to the groups	RIAD reference data dataset
Size of the group	The group size calculated as described in section 3	Author's calculations
Significance multiplier	The baseline significance multiplier based on pure counts for each group member	Author's calculations
Business specific significance multiplier	The business specific significance multiplier based on the balance sheet value for each group member	Author's calculations

References

Schnabel, I. (2020): Don’t take it for granted: the value of high-quality data and statistics for the ECB’s policymaking

Tissot, B (2019): Financial big data and policy work: opportunities and challenges

Tissot, B and De Beer, B (2020): Implications of COVID 19 for official statistics: A central banking perspective

Tissot, B. and De Beer, B. (2021): Enhancing the global architecture supporting economic and financial statistics: a post Covid19 perspective

Rosolia, A., Stapel-Weber, S., Tissot, B. (2021): New developments in official statistics

Romano, S., Martinez-Heras, J., Raponi, F., Guidi, G., Gottron, T. (2021): Discovering new plausibility checks for supervisory data

EUROPEAN CENTRAL BANK
EUROSYSTEM

Bruno Carreiras
DG-Statistics/AMA/MAM, European Central Bank

Data quality management of entity group data

Relevance and tools to address current challenges

IFC 11th Biennial Conference

Post-pandemic landscape for central bank statistics

25-26 August 2022, Basel

Overview

Context

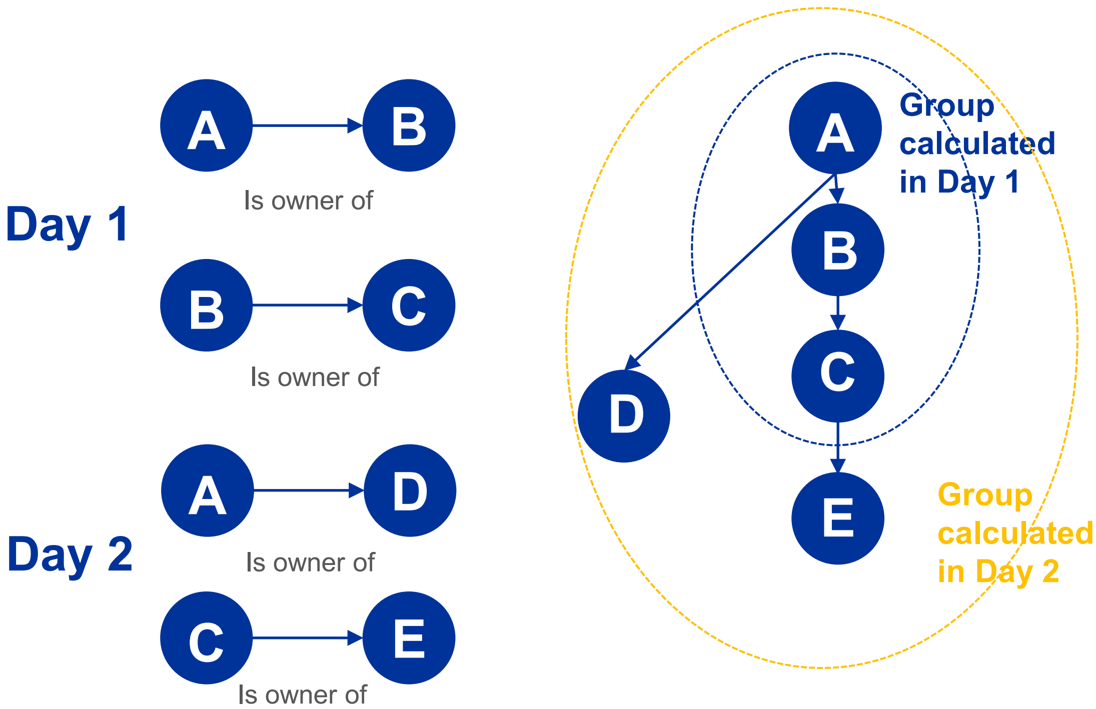
- The pandemic increased the need to **deliver data** to decision-makers **quickly and with quality**.
- Managing these expectations is **challenging**, especially when dealing with **large granular datasets** as the Register of Institutions and Affiliates Data (RIAD).
- Being the ESCB shared entity master-dataset, RIAD is managed within a **highly dynamic environment** where attributes and relationship data can be changed every day.
- Very large datasets combined with regular changes will necessarily need measures to **monitor and ensure data quality efficiently:**

> This paper tries to respond to that need by presenting a metric, the **Significance Multiplier**, applied to the RIAD Group Networks that aims at **prioritising potential data quality issues** by evaluating movements in group structures.

How are groups created in RIAD?

In RIAD, **relationship data is collected from NCBs and NCAs** (based on national data sources) and disseminated to users **daily.**

– Based on such relationships, **group network trees are <u>calculated</u> each day (dynamic)**

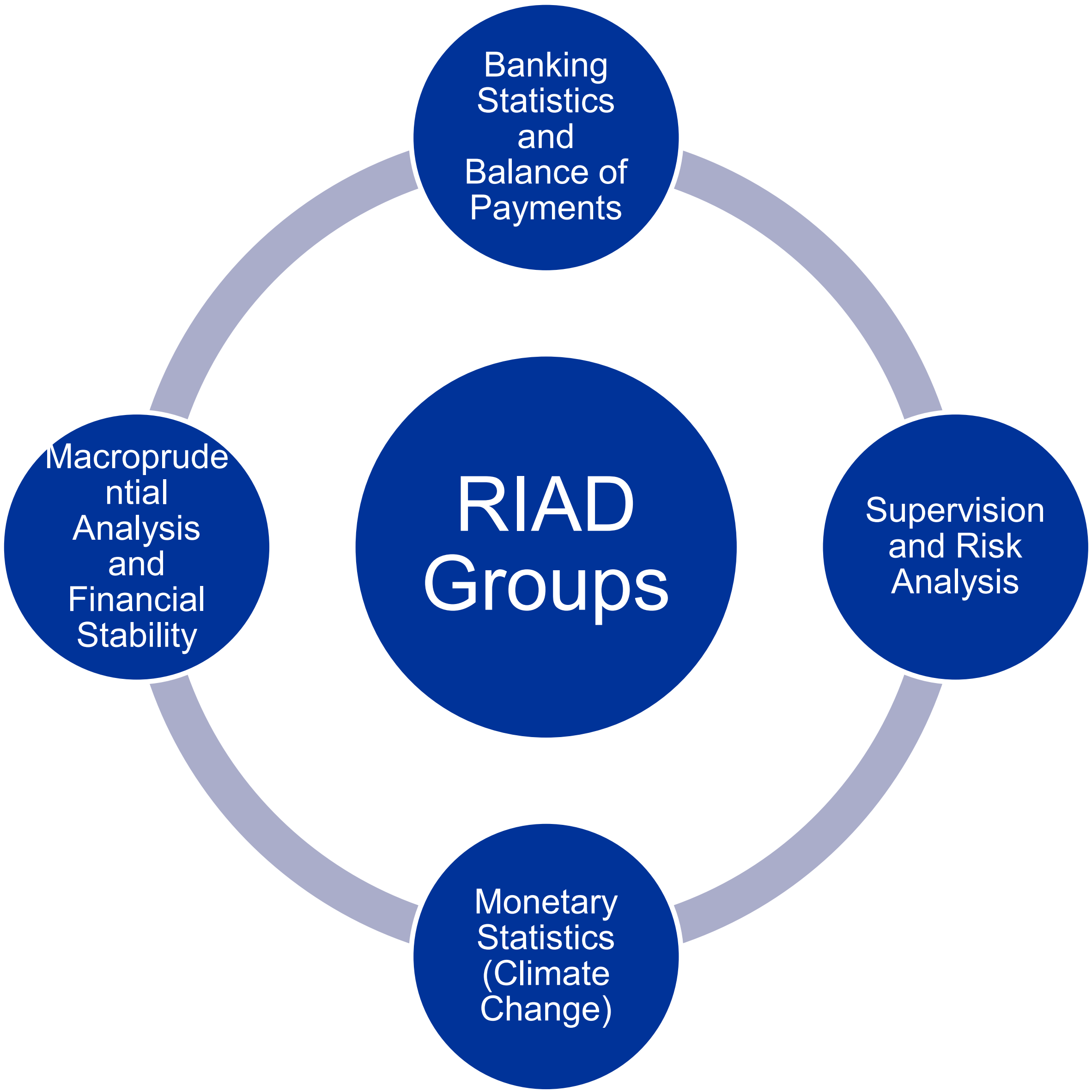


~300,000 ownership relationships

~130,000 groups

Available to users via a **large network dataset**

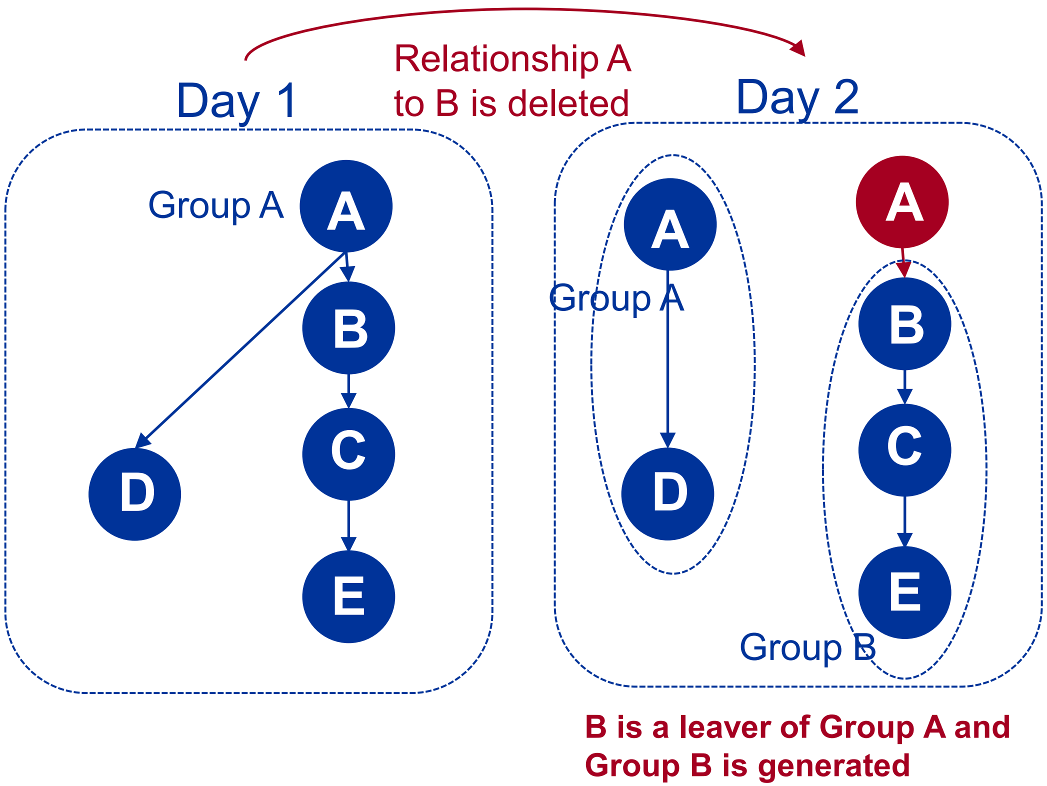
Annex – Users of RIAD Groups



Ensuring groups' data quality: group movements

DQM for groups is performed in two ways:

- Consistency checks on the relationship data (pre-calculation)
- **Monitoring <u>movements</u> in group structures: real business changes vs erroneous changes (post-calculation)**



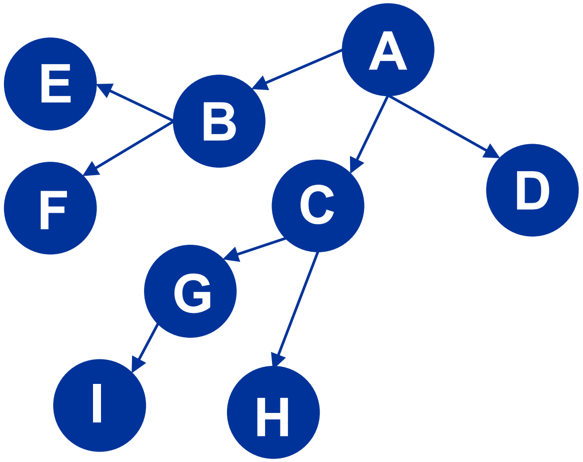
Possible mistakes impact users → **movements need to be assessed**

In a universe of 120,000 groups that may change every day, it is important to **<u>evaluate and prioritise</u> the monitoring and verification of movements**

Finding the most relevant movements

In the context of the RIAD groups: (1) movements always start in a single event on a group member and (2) if a member leaves it takes all its direct and indirect children.

At the baseline, the relevance of a group member can be given by the count of members affected by an event on that member or, **Significance Multiplier** → the size of **(sub)group headed by the affected entity**.



Group member	Significance Multiplier
A	9
B	3
C	4
D	1
E	1
F	1
G	2
H	1
I	1

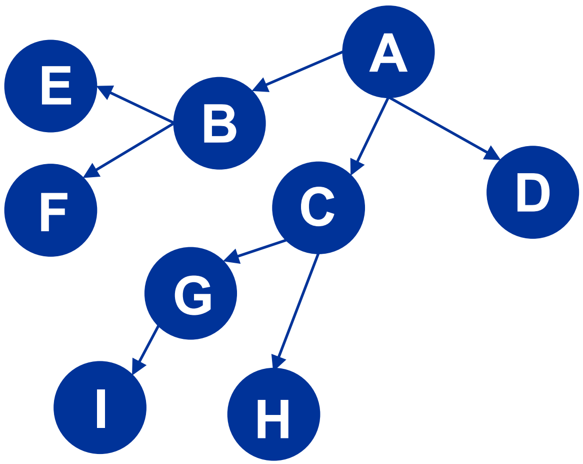
Measure of relevance for group member and movements intra and cross group

Larger groups and members with more children are always considered more relevant

What is relevant for you?

RIAD Groups are used by multiple stakeholders → important to account for their priorities and needs in any data quality procedure and related DQM prioritisation.

- Extension: the **valuation should be based on a criterion selected by the interested business area (α)** rather than a pure count of the movements. The Significance Multiplier stops being the count of all the members of a subgroup but the sum of its α (αSM)



Group member	SM	αSM
A	9	$\alpha_A + \alpha_{B,\dots,I}$
B	3	$\alpha_B + \alpha_E + \alpha_F$
C	4	$\alpha_C + \alpha_G + \alpha_H + \alpha_I$
D	1	α_D
E	1	α_E
F	1	α_F
G	2	$\alpha_G + \alpha_I$
H	1	α_H
I	1	α_I

Prioritisation ranking for data quality is based on a defined criterion and not on a naïve count

By-product: Users can **evaluate every group member** for their business purpose (can be calculated daily)

Defining a priority ranking for DQ

SM for priority ranking → relevance of groups and group members gives an indication on which movements should be assessed first.

By group

Group Head	SM	αSM (Bln €)	SM Ranking	αSM Ranking
Banking group 5	696	>2,000	5	1
Banking group 7	363	>2,000	7	2
Banking group 8	348	>2,000	8	3
Banking group 17	137	>1,000	17	4
Banking group 15	186	>1,000	15	5
Banking group 6	348	>500	9	6
…	…	...	…	…
Banking group 233	50	>100	36	12

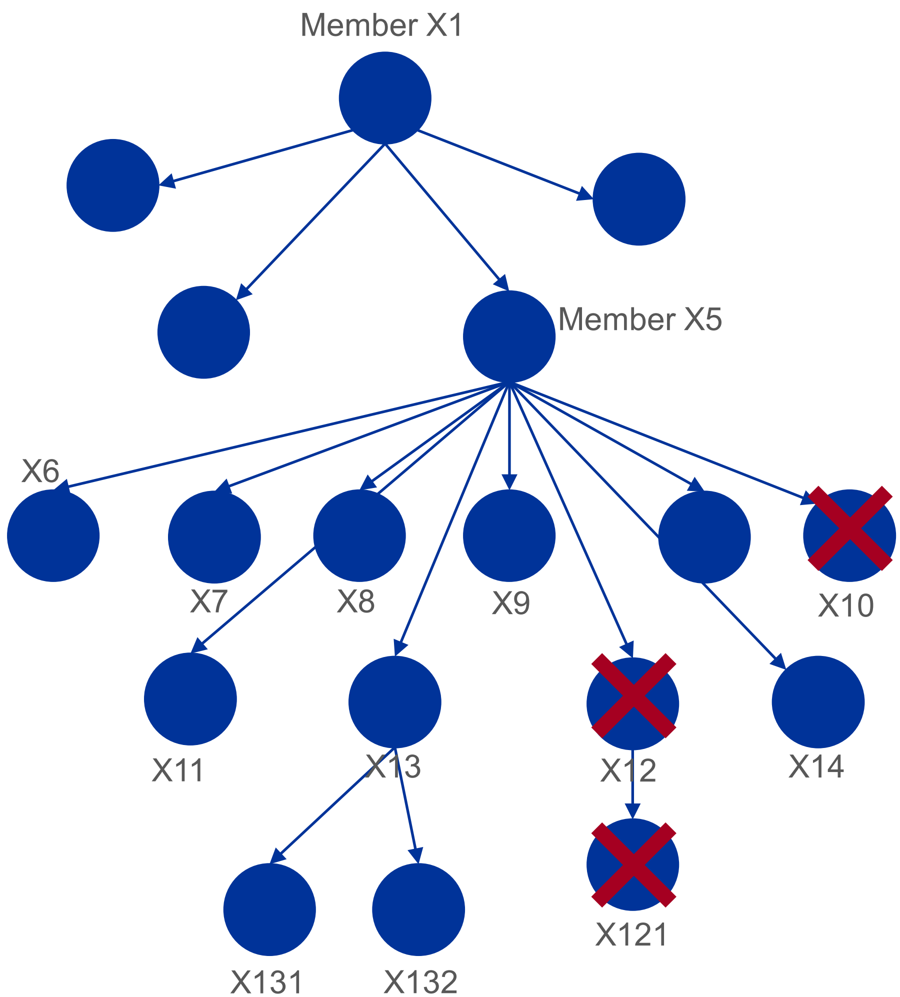
By group member

Group Member	SM	αSM (Bln €)	SM Ranking	αSM Ranking
Member 320	21	>2,000	>120	1
Member 11	194	>2,000	11	2
Member 321	2	>2,000	>500	3
Member 323	1	>2,000	>500	4
Member 324	25	>1,000	>120	5
Member 325	5	>1,000	>500	6
…	…	...	…	…

Movements are given the right priority!
→ movements on **relevant branches are assessed first**

Applied to: RIAD Banking groups, end-May
Business specific criteria (α): Individual Balance Sheet

Valuation of corporate changes – the Banking Group X



X12 and X10 branches are closed in RIAD reflecting real world changes!

What is the **impact of each of the changes** on banking group X balance sheet?

$$\alpha SM_{X12} = \alpha_{X12,Jan22} + \alpha_{X121,Jan22} > 1{,}500\ Mln\ €$$

$$>$$

$$\alpha SM_{X10} = \alpha_{X10,Jan22} < 1{,}500\ Mln\ €$$

Changes are ranked by relevance intra-group!

Valuation is already available before the event (Jan)!

Applied to: RIAD data on banking group X, end-Jan vs end-May
Business specific criteria (α): Individual Balance Sheet

Conclusions

- The **Significance Multiplier** provides benefits in two ways:
  - Provides a **prioritisation measure** for data quality management of changes in group networks;
  - Provides an **evaluation technique** for corporate changes for assessing the impact of movements in *real time*.
- Already being used to prioritise data quality work on RIAD group networks.
- **Possible extensions:**
  - Application to other business cases: climate change (carbon intensity) or credit data;
  - Application to other network datasets under the same business conditions (dynamic environment).

Thank you! Questions?

Bruno.Carreiras@ecb.Europa.eu