A small step for EMIR, a giant leap for transparency – first steps towards identifying strategies for boosting data quality in granular data on derivatives

Gemma Agostoni[1], Nicola Calabrese[2], Marco D'Errico[3], Lukas Henkel[4], Grzegorz Skrzypczynski[5]

[1] European Central Bank, Directorate General Statistics

[2] European Central Bank, Directorate General Statistics

[3]European Systemic Risk Board Secretariat

[4] European Central Bank, Directorate General Statistics

[5] European Central Bank, Directorate General Statistics

Abstract

The European Market Infrastructure Regulation (EMIR), a significant piece of the EU post-crisis legislation, introduced several requirements for EU-based counterparties entering derivative contracts, including the obligation to report granular transaction-level information daily to Trade Repositories (TRs). This information is then shared with over one hundred EU authorities.

The ECB and the ESRB staff have been intensively using this data, leveraging a dedicated, high-performance IT infrastructure. The information has been used to work on a set of diverse tasks, facilitating timely monitoring and analyses of derivative markets, progressively becoming a building block for risk analysis and policy development and assessment.

However, nearly a decade since mandating the reporting requirements, the biggest challenge in exploiting the data are the persistent and significant data quality problems. These issues do not lie within the reporting framework effort but to the limited commitment by reporting entities (including central counterparties and large financial institutions) to developing internal processes for validating and checking the quality of data before submitting them to the authorities.[1] While there has been some improvement over years, data reported under EMIR continues to fall short of being fully satisfactory for scaling and automation. Despite employing advanced IT tools and having accumulated a vast expertise, it still requires considerable time and effort to clean, aggregate and analyse these data. Given these persistent problems, conclusions drawn from the data can carry considerable uncertainty: is a specific pattern indicative of a relevant financial development or merely a by-product of poor data quality?

[1]See: European Systemic Risk Board (2024)

This paper describes how EMIR data is used at the ECB and the ESRB Secretariat, highlighting the primary data quality challenges the experts encounter when working with the data. Our main contribution is to delineate the most common and significant data quality issues inherent to EMIR in order to propose a comprehensive taxonomy. This classification enables us not only to identify the root causes of these issues, but also explore the reasons for their persistence over time. To our knowledge, this is the first paper to attempt such classification.

We build on this work to offer recommendations for enhancing data quality. Although it is unrealistic to anticipate the immediate resolution of all issues in a dataset of this scale and complexity, we propose that the implementation of certain systematic and potentially automated solutions could significantly improve the quality of the data. Remarkably, this improvement could be achieved with minimal effort from reporting entities and regulators.

Finally, we argue that enhanced data quality will boost the use of this information by regulators and policy makers. This progression would facilitate market monitoring, leading to further data quality improvement. Even more importantly, this would bolster financial stability of the EU and its member states. We also believe that streamlined procedures and better data quality management could benefit the industry itself.

Keywords: EMIR, derivatives, data quality, granular data, regulatory reporting

1. Introduction

The European Market Infrastructure Regulation (EMIR) is an EU law that aims at ensuring well-functioning and transparency of derivatives markets in Europe. The legislation entered into force in 2012 and was a response to the commitment of G20 leaders in September 2009 at the Pittsburgh summit to improve the over-the-counter derivative markets.[2] One of the three main pillars of the Regulation is the obligation to report granular transaction-level information on derivative contracts to Trade Repositories (TRs), specialised entities, authorised by the European Securities and Markets Authority (ESMA). The data is then made available to over one hundred authorities in the EU.

The ECB and the ESRB Secretariat have been actively using these data for various tasks, including financial stability monitoring, micro-prudential supervision, and Central Counterparties' (CCPs) oversight. However, since the inception of reporting, the information collected under EMIR suffers from severe data quality issues. Even though the situation has improved since 2014, the data is still extremely difficult to be used effectively after nearly a decade of reporting obligation, requiring considerable manual effort to clean, meaningfully aggregate and analyse the data.

The recognition of insufficient quality as one of the main limits in the efficient use of data on derivatives to exercise authorities' tasks has been gaining traction among regulators. The

[2] See: https://www.g20.utoronto.ca/2009/2009communique0925.html

recent compromise text[3] for EMIR amendment, so called EMIR 3.0, notes: “Entities subject to the reporting obligation pursuant to that Regulation should therefore be required to exercise due diligence by establishing appropriate procedures and arrangements to ensure data quality before submitting their data.” Also, the industry notices the need to improve processes related to risk managements by increasing automation, interoperability and data standardisation.[4]

This paper briefly describes the experience of the ECB and the ESRB Secretariat with data collected under EMIR, mainly through the lens of data quality. We attempt to identify and classify the most frequent and prominent issues encountered when working on these data. Building on this classification, we offer initial recommendations on how data quality could be enhanced, with a few conceptually simple measures. While their full implementation might require a certain evolution of mindset of the regulatory community, we believe it would greatly enhance the usefulness of this information by regulators. Better quality would subsequently reduce “barriers of entry” for potential new users in the regulatory community, and lead to even better scrutiny and, finally, to more robust and stable derivative markets in the EU.

The paper is organised as follows. Chapter 2 briefly sketches the development of data quality in EMIR over time, and data quality assurance processes at the ECB and the ESRB Secretariat. It also presents the classification of types of data quality problems encountered in EMIR. Chapter 3 presents a few recommendations aiming at improving the quality of data. Chapter 4 concludes.

2. Data quality in EMIR

2.1 Evolution of data quality in EMIR

The reporting obligation under EMIR started in February 2014. Since then, authorities in the EU got access to an unprecedented wealth of data on derivative transactions traded in Europe. From the beginning, however, these frequent, granular data were plagued by quality issues, limiting the regulators in the ways they can use this information.

Under EMIR, data are reported by market participants to trade repositories (TRs) – dedicated entities authorised by ESMA, which are then obliged to compile and submit the respective reports with granular data to authorities. Initially these reports from TRs suffered from a lack of standardisation: various channels and formats were used to transmit information collected under EMIR, with significant differences among the TRs. Hardly any validation was carried out

[3] See: https://data.consilium.europa.eu/doc/document/ST-6344-2024-INIT/en/pdf

[4] See: https://www.isda.org/2023/02/20/automation-and-data-standards-in-collateral/

on the input data provided to TRs, leading to serious quality issues and missing information passed to the regulators. These issues made it extremely difficult to aggregate and analyse information by the authorities.

Since then, EMIR data improved notably, due to various measures implemented by ESMA, EU Commission, and other EU authorities. In particular:

- In 2014 and 2015, ESMA introduced mandatory validation rules that TRs are obliged to carry out on the data they receive from market participants;
- ESMA obliged the TRs to carry out a weekly reconciliation exercise, where two legs of individual trades are being identified within or across TRs, and subsequently compared for a subset of reported attributes;
- ESMA introduced EMIR Questions & Answers – regularly updated guidelines for market participants and TRs;
- In November 2017 a major revision of technical standards for reporting introduced new fields and common format for transmission of data from TRs to regulators, compliant with the ISO20022 standard;
- Continued dialogue and sharing of experiences between authorities within and outside EU increased the expertise on the data among the users;
- The ESMA data quality framework enabled authorities to report data issues to ESMA, for their prioritisation and distribution to competent national supervisory authorities.

Although these measures led to considerable improvements, the quality of EMIR is still one of the biggest obstacles in the efficient use of this information for authorities' tasks. The validation rules will not cover all possible cases of misreporting, and even the most accurate guidance may be ignored or misunderstood by the reporting entities. ESMA's data quality framework is an essential tool for coordinating the supervisory response to quality issues on the EU level, but the complexity of the framework (27 national supervisory authorities, ESMA supervising the TRs, and over 100 authorities using the EMIR data) leads to bottlenecks and long feedback loops in the data quality assurance process.

Consequently, these data still require significant effort to understand, clean, and meaningfully aggregate. The following sections describe main challenges and how this problem is dealt with at the ECB and ESRB Secretariat.

2.2 Data quality assessment at the ECB and ESRB Secretariat

In their analysis of EMIR data, the ECB and the ESRB Secretariat rely on the ECB EMIR IT system, a dedicated, modern IT infrastructure to collect, process and analyse the tens of

millions of transactions reported daily under EMIR. The IT system went live in 2017, and greatly facilitated analysis of this rich dataset.

On top of that, the ECB and the ESRB Secretariat developed a range of tools and processes to identify, analyse and correct for quality issues present in the data. First, a daily monitoring process, supported by a daily automated report, is carried out by the responsible team, to verify completeness of the collection, as well as correctness of data ingestion, enrichment from reference data sources, data deduplication,[5] and presence of outliers.

The next layer of data quality assurance consists of more sophisticated analysis of data developments. For this purpose, ECB and ESRB Secretariat staff built an Automated Data Quality (ADQ) framework, a customizable, automated tool based on decision trees, aiming at identifying and classifying the developments of numerical variables, like notional amounts or contract values.[6] The results are analysed daily and logged in a dedicated tool, where they are available for the broad community of EMIR users at the ECB and ESRB Secretariat.

Finally, a set of cleaning procedures was designed to make data ready for analytical purposes. The procedure performs a series of operations, identifying data quality shortcomings and taking measures to correct them.[7] The outcome is a new table that can feed into further analysis of the data.

The data quality of EMIR benefits from the fact that it is a double-sided regime (meaning that the same trade is reported twice, if both counterparties are located in the EU) and also from the comparison with other data sources. The trade details of paired legs (i.e. trades that are reported by both counterparties) can be compared allowing to detect inconsistencies that should be then further investigated. In addition, EMIR partially overlaps with the data reported under the Money Market Statistical Regulation (MMSR)[8]. This is limited to the euro overnight index swaps market and the FX swaps denominated in euro. Procedures to compare this information at macro and micro level have been developed leveraging not only on the unique trade ID but also on algorithms that allow to pair and match the same trade in the two datasets.

---

[5] Reporting under EMIR is double-sided, meaning that both counterparties are obliged to report details of the trade. In order to avoid double-counting of trades when aggregating, data has to be de-duplicated.

[6] See Agostoni, et al. (2023)

[7] See Agostoni, et al. (2024)

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

2.3 Taxonomy of EMIR data quality issues

As described in the previous chapter, data collected under EMIR suffer from considerable problems with data quality, which make the usage of these data challenging. In this section, we classify the data quality issues we have been encountering into categories. Though this list is by no means exhaustive, and will be expanded through further investigation, in our view it captures the most prominent challenges with EMIR quality. In the following sections we describe each category providing its main significant features.

2.3.1 Implausible (large) numerical values

Implausible values for numerical variables in EMIR, like contract and notional values of trades or margins posted and received by counterparties, are a primary concern for data quality and economic analysis of EMIR data. It is crucial to understand whether a large change in, e.g., the total contract value of outstanding contracts of a specific entity, indicates a genuine economic development, with possible consequences for financial stability, or this change is an artifact due to misreporting of possibly even a single trade.

Since even statistical outliers might reflect genuine behaviour of reporting entities, a purely statistical approach is bound to misidentify genuine reporting of trades as outliers in some cases. Hence, identifying implausible numerical values in EMIR data needs two components:

1. Detection of statistical outliers
2. Expert judgement whether a statistical outlier reflects a misreporting behaviour.

While the detection of statistical outliers can be performed in an algorithmic manner, the decision whether an outlier indeed is implausible is challenging to automate, given the complexity of EMIR data and reporting framework and usually requires expert judgement. This is further complicated by the fact that the implausible numerical values can appear starting from the initial reporting of a trade, or reported values can change by several orders of magnitude when information on a trade is updated. Hence, even if the initial reporting of a trade was correct, implausible values might be reported later in the lifecycle of a trade. An abnormal change in a numerical variable might also reflect the correction of previously misreported values.

To identify implausible numerical values in EMIR data, the ECB and the ESRB Secretariat have developed the ADQ framework which consists in an elaborated algorithm based on decision trees[9].

[9] See Agostoni, et al. (2023)

Implausible numerical values, regularly found for the main numerical variables in EMIR such as contract values, notional amounts, and margins, can be grouped into different categories:

- **Idiosyncratic trade errors.** A single value is misreported for a single trade, reported by an entity that otherwise reports no implausible values. These errors are usually corrected within a few days but might persist longer.
- **Systematic errors with correction.** Specific entities regularly report implausible values for the same variable(s), for trades with similar characteristics. After the initial reporting of implausible values, the trades are regularly corrected to plausible values within a few days after the initial reporting.
- **Systematic errors without correction.** Similar to the previous case, specific entities regularly report implausible values for the same variable(s), for trades with similar characteristics. In contrast to the previous case, the implausible values are not corrected over the lifecycle of the trade, or only for a small subset of trades with the relevant characteristics.[10]
- **Chaotic entities.** A small subset of entities regularly reports implausible numerical values for many trades, but the trades do not share common characteristics except the reporting entity. Usually, in addition to reporting implausible values, the numerical reported values exhibit significant and unusual volatility at the trade level for these entities.

The implausible numerical values pose a significant problem for the users of the dataset, as these items have to be cleaned, often in a laborious, manual process. The task is significantly easier if the issues follow a predictable pattern. The challenge grows if the pattern is irregular. For this reason, especially the fourth type of implausible numerical values is an issue for users of EMIR data.

2.3.2 Reporting of margins

Margin requirements are at the heart of the regulatory reforms concerning derivatives. Derivative instruments transfer risk between counterparties, with their value dependent on the stochastic cash flows they generate over time, encapsulating both underlying risk and counterparty risk. Counterparty risk is mitigated through the exchange of initial and variation margins, which can introduce funding liquidity risks if counterparts lack the immediate funds to

[10] We can also observe that certain types of misreporting are more prevalent for specific asset classes or products, which may hint at not sufficiently developed reporting guidance or differing market practices. An example are contracts for difference (CFD), where often the reported contract value is close to the notional amount and is clearly too high to reflect the replacement value of the contract.

meet margin calls. This scenario can force counterparts to sell assets quickly, creating market liquidity risks through potential "fire sales" or rapid closing of derivative positions. Thus, valuation, margin requirements, and the interconnected nature of these risks are essential in understanding financial stability.

Regulatory data on derivatives must therefore quantitatively assess and integrate underlying, counterparty, funding liquidity, and market liquidity risks, ensuring that the data is coherent from a risk perspective. Such coherence is crucial for regulators to effectively understand and mitigate systemic risks associated with derivatives. This coherence also reflects data quality, evaluated by consistency in risk measurement, analysis of risk interrelationships, and the predictive validity of data. The interaction between contract valuation and variation margins illustrates the quantitative relationship among identified risk types. Variation margins should align with historical fluctuations in underlying risk and are quantified through financial models or market values. Understanding this relationship enhances risk assessment accuracy and the robustness of the risk monitoring framework.

Despite their centrality in the post crisis regulatory framework, and their centrality in the risk management of institutions, margin data has suffered from severe data quality issues which have hampered policymakers' ability to monitor risks.[11] The European Systemic Risk Board (ESRB) has highlighted significant concerns regarding data quality limitations in CCPs.[12] These limitations compromise the primary objective of enhancing transparency in financial transactions, directly contravening the foundational principles that underpin the advocacy for central clearing mechanisms. This issue underscores a critical challenge in achieving the desired transparency for CCPs.

Data quality issues with margin data can be grouped into the following categories:

- **Lack of uniformity in reporting.** Despite the well-defined nature of the reporting rules, entities frequently adopt arbitrary criteria and modify them over time. This variability undermines the capacity of authorities to enhance their monitoring activities, necessitating periodic manual interventions to correct these inconsistencies.

[11] See, for instance, the ESRB's letter to EU co-legislators on the EMIR review (available here: https://www.esrb.europa.eu/pub/pdf/other/esrb.letter230320_on_emir_review_mep~058e272ec7.en.pdf?406179830229e8e1aa32068c52f22f7b) and the ESRB response to BCBS, CPMI, IOSCO (European Systemic Risk Board (2024))

[12] See: European Systemic Risk Board (2024).

- **Lack of standardisation.** Collateral data lacks standardisation for both internal and external reporting and does not adhere to basic internationally agreed standards, such as the LEI (Legal Entity Identifier), UTI (Unique Transaction Identifier), and UPI (Unique Product Identifier). For example, there is no such concept as a UMPI (Unique Margin Portfolio Identifier), which would enable entities and authorities to unequivocally map which trades are included in a specific portfolio.
- **Lack of reconciliation.** Collateral data cannot be reconciled not only due to the absence of standardised formats and identifiers, but also because entities fail to perform reconciliations regularly, since it is not mandated. Furthermore, it is virtually impossible to determine the appropriate level of aggregation for such reconciliations.
- **Implausible values.** Some entities, including CCPs, report values that are clearly inconsistent with market valuations. These discrepancies can be significant, sometimes spanning several orders of magnitude.

2.3.3 Reporting the direction of exposure

In cases where both counterparties of a trade are obliged to report, when reporting contract values the two counterparties are expected to report an opposite sign, i.e. if counterparty A reports trade X with counterparty B and contract value +100, then the counterparty B should report -100.[13] However, for a large number of cases this convention is not followed, and both counterparties report the contract value with the same sign (usually both positive, but there are also cases where both counterparties report negative contract values). This poses a serious challenge in the analysis of derivative exposures.

The reason for the above phenomenon may be often traced to the fact that the valuation is done by the more sophisticated counterparty (e.g. CCP, clearing member, financial company) and then passed to the less sophisticated counterparty (e.g. non-financial company), which applies the valuation without understanding the need to reverse the sign of the contract value. Nevertheless, this is rather a general observation, which may be difficult to apply to all possible scenarios – highlighting the need for particular scrutiny, when analysing contract values in EMIR.

[13] Slight differences in the absolute value are possible, due to differing valuation date/times, and/or different valuation models, if the derivative is not marked to market.

2.3.4 Errors in reporting trading lifecycle

In EMIR the counterparties are obliged to report not only new trades, but also events affecting the trade during its lifecycle. These so-called “lifecycle events” can take the following values in EMIR reporting:[14]

- New
- Modify
- Error
- Early termination
- Correction
- Compression
- Valuation update
- Position component

Based on the values reported, the TRs determine the status of a derivative contract for a given date and use this information to populate the trade state report.[15] Hence, the correctness of information provided to authorities on the outstanding state of derivatives relies to a large extent on the lifecycle information provided by the counterparties.

The pre-Refit EMIR framework suffered from one important flaw: apart from valuation update and early termination, the reporting template does not allow the counterparty to specify at what time an event happened. With a few exceptions, the only available information is the timestamp specifying when the update was submitted to the TR, which does not necessarily reflect the moment when the event actually took place.[16] The TRs are not able, thus, to chronologically order the lifecycle events reported for a certain trade and they have to make certain assumptions when building the trade state reports, e.g. handling the lifecycle events in the order in which they were reported. This could result in a wrong interpretation of the lifecycle of the contract.

To illustrate this issue with an example, let us consider a sequence of reporting by a specific counterparty for an individual trade:

---

[14] For definitions, see Commission Delegated Regulation 148/2013 (as amended), Annex, Table 2, row 93. Please note that these categories changed under EMIR Refit new reporting rules (see section 3.1)

[15] Based on activity data received from their clients, TRs are obliged to compile the so-called “trade state report”. The report includes the most up-to-date state of all outstanding derivative contracts.

[16] Under EMIR, the entities are obliged to provide details on the lifecycle events on the day following the event, at the latest. This obligation, however, is often not satisfied (especially for corrections).

Table 1: Example of reporting sequence

Date of reporting to a TR	Action type	Actual date of the event (not reported)
2024-04-02	New	2024-04-02
2024-04-10	Early termination	2024-04-09
2024-04-12	Correction	2024-04-05

As presented in Table 1, the TR would start including the trade in the trade state report on 2024-04-02, when the trade is reported as New, and remove it on 2024-04-10 as result of the reporting of an early termination. However, the TR cannot unambiguously determine the meaning of the last event reported, i.e. the correction. The TRs employ different strategies for building the trade state reports: (i) they may decide that the entity wanted to withdraw the early termination and put the trade back in the trade state report; (ii) alternatively, they may treat the third message as a mistake and reject it. Without the proper knowledge of the underlying situation, a valid trade may not be included, or an expired trade may be reintroduced into the report.

Another set of lifecycle-related data quality issues is linked to the position-level reporting. EMIR allows for reporting of derivatives on either transaction- or position-level. The latter means that individual trades in fungible products are replaced by a position, as of when the counterparty has to report lifecycle events only for the position, not the individual trades. Some counterparties do not strictly hold to the transaction/position distinction and report different lifecycle events for a given trade as transactions or positions interchangeably. This creates difficulties for TRs, similar to the ones described above. If a new trade is reported as a position, but then the valuation update is provided on the transaction level, it is not obvious whether this is a valid update or a different trade with erroneously reported identical trade ID.

The above problems hamper the accuracy of the trade state report, which is the main report used by the regulatory community. The EMIR Refit amendments, however, extend the EMIR lifecycle model, making the information to be reported more detailed and accurate – see section **Erro! A origem da referência não foi encontrada.** for details.

2.3.5 Errors resulting from delegation of reporting

The EMIR Regulation foresees that, in some specific cases, an entity that is not the counterparty is legally liable for submitting the report. In particular:

- A financial counterparty is responsible for reporting details of an OTC derivative transactions on behalf of a non-financial counterparty under the clearing threshold (unless the non-financial counterparty waives this right);

- A management company is responsible for reporting details of an OTC derivative transaction on behalf of a UCITS investment fund;
- An alternative Investment Funds' Manager (AIFM) is responsible for reporting details of an OTC derivative transaction on behalf of an Alternative Investment Fund (AIF);
- If an Institution for Occupational Retirement Provision (IORP) does not have legal personality, the authorised entity that is responsible for managing and acting on its behalf is responsible for reporting details of an OTC derivative transaction.

The entity can also voluntarily delegate the reporting obligation to their counterparty, or to a third party.[17]

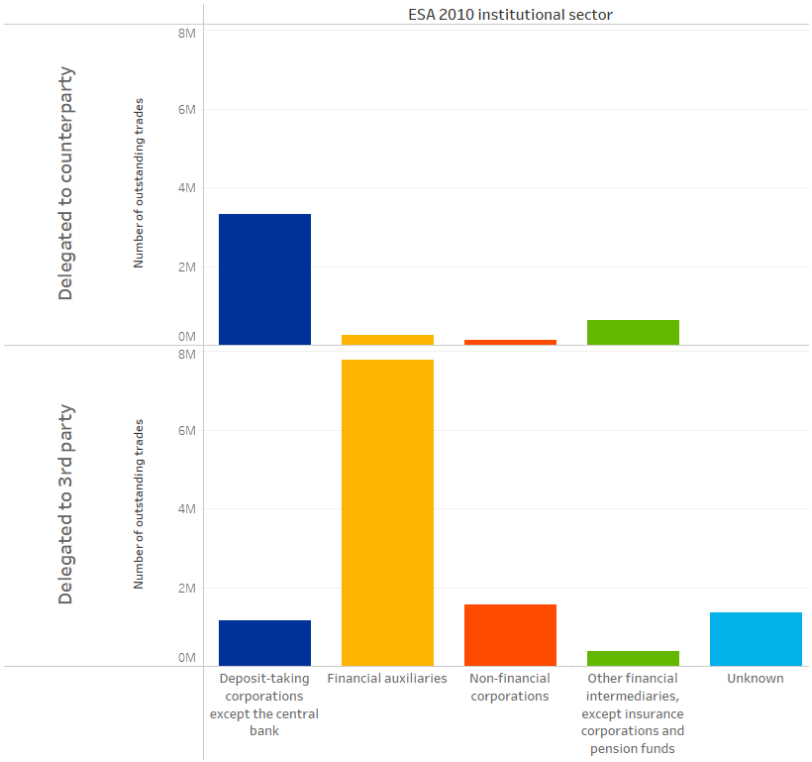
Delegation of reporting is quite prevalent in EMIR, with around 15% of outstanding trades delegated to the other counterparty of the trade, and over 40% delegated to a third-party. Figure 1 below presents the ESA 2010 sector[18] of the entity submitting the report, conditional on the delegation type. If the reporting is delegated to the other counterparty, this counterparty is usually a bank.[19] In case of third-party delegation, the composition is a bit more diverse, but the financial auxiliaries are clearly on the lead. This group includes different entities, including brokers, trading venues, and financial services companies providing regulatory reporting services.

[17] See Article 9(1a) of Regulation (EU) No 648/2012 (as amended)

[18] See: https://ec.europa.eu/eurostat/web/esa-2010

[19] It is also worth noting that for over 30% of the trades delegated to the other counterparty, the latter is an entity from the same group.

Figure 1: ESA sector of entities to which the reporting was delegated in EMIR.



Source: ECB EMIR data and ECB RIAD (Register of Institutions and Affiliates Data). Not relevant categories are not included.

One of the data quality issues stemming from reporting delegation is lack of agreement between counterparties on who should submit the report. In EMIR we can observe cases where the lifecycle event is submitted by the reporting counterparty, but at the same time, the same lifecycle event is submitted by the entity to which the reporting was delegated. This may result in two types of issues:

- the two messages are reported to different TRs → in the data available to the regulators, the same trade appears twice;
- the two messages are reported to the same TR → the TR updates the information from two different sources in alternating manner; information received later during the day will become end-of-day state of derivative on trade states report.

The former case presents a clear challenge for the regulators, as the information has to be cleaned of duplicates. The issue concerns around 13,000 trades outstanding at a given moment of time.

The latter scenario seems innocuous at first glance, as it does not lead to duplicates in the data received by the regulators. The problem arises, however, when the information reported

by the two entities submitting the message is not consistent. This leads to the details of the outstanding trade alternating randomly between two sources of information. For instance, we observe cases where the contract value of a trade jumps back and forth between two very different values, following an irregular pattern. Such problems are difficult to spot, interpret, and correct by the researcher analysing the data.

2.3.6 Timestamps and time zones

According to the EMIR regulation, timestamps are supposed to be reported as UTC timestamps. However, not all reports adhere to this, and might report their local time, rather than UTC timestamps. This can, for example, lead to the issue that prices reported by an entity differ significantly from exchange-quoted prices for exchange-traded derivatives.

Another issue related to timestamps are missing timestamps (e.g. missing maturity dates), or implausible timestamps.

2.3.7 Bugs in TRs' IT system

As described in section 2.1, TRs constitute an intermediate layer between data reporters (market participants entering into derivative contracts) and data users (EU authorities). This process is not a simple pass-through – TRs are obliged to verify the correctness of the reported information and compile reports for authorities. The latter process presents some implementation challenges, given that:

- The output and input format are different;[20]
- The TRs have to compile dedicated reports – in particular trade state reports, where the last state of outstanding derivatives have to be inferred from lifecycle events reported to the TR;
- The scope of the reports depend on mandates of the receiving authorities – hence the TRs have to filter transactions per each authority.

In such setting, the TRs have to carefully design the pipeline to "translate" the input to output reports. Any mistake in this process leads to errors in the reports received by the regulators, which may include:

- Wrong scope, due to incorrect filtering (too many or too few transactions),
- Incorrect representation of the outstanding state of the contract,

[20] Before EMIR Refit, TRs implemented proprietary input formats, while they were obliged to transmit the data to the regulators using a standardized ISO20022 message. This problem is expected to be alleviated with EMIR Refit – see section **Erro! A origem da referência não foi encontrada.**.

- Missing transactions or superfluous transactions, due to incorrectly applied validations (too strict or too lax),
- Data lost or tampered in the course of processing.

In the 10 years of EMIR reporting, the ECB, ESRB Secretariat, and other EU authorities identified multiple problems of this type. Most of those items have been fixed, in a collaborative process involving authorities, ESMA and TRs.

2.3.8 Underreporting

One of the features of EMIR is the double-sided reporting obligation, which means that, unless the second counterparty to the derivative is domiciled in a non-EEA country, the details of a trade between two EU entities must be reported for each of the counterparties.

Underreporting in EMIR data occurs when one side of a transaction fails to report necessary details. Despite regulations mandating both parties to report the same trade separately, challenges in ensuring full compliance persist. This results in discrepancies between reported derivatives, making it difficult for authorities to accurately assess entities' exposures. For instance, in 2018, four years after EMIR implementation, up to 55% of trades remained unpaired, meaning one counterparty reports a transaction while the corresponding report from the other counterparty is missing.[21] Out of these unpaired trades, around 55% could be justified by one of the two reasons:

- Trades with entities from non-EU jurisdictions: EMIR applies to entities resident in the EU, thus the trades concluded with counterparties outside the EU are expected to appear only once in the dataset.
- Trades with natural persons: The EMIR regulation does not impose the reporting obligation on natural persons, hence for trades carried out by private individuals we will see only the leg reported by the counterparty of the trade.

The remaining 45% non-paired trades seemed to be a case of misreporting, caused by:

- Misreporting of counterparties' IDs: the reporting entity submits a valid LEI code as the ID of the other entity, but this is not the LEI with which the other counterparty identifies itself, thus the two trades remain unpaired.
- Non-outstanding trades on trade state report: one counterparty reports expiration of a trade, but the other one fails to do so – in this case one leg of the expired trade remains on the trade state report and cannot be paired.

[21] See: Pérez-Duarte, Skrzypczynski (2018)

- Underreporting: one of the counterparties failed to meet its reporting obligation.

The number of non-paired trades that are considered misreporting decreased when the UK left the European Union, but this merely reflects the fact that we do not expected to pair trades executed with UK counterparties anymore. The issue of non-paired trades remains a significant problem, with around 2 million trades unpaired without a legitimate reason.[22]

Furthermore, in the process of providing data to the authorities, TRs implement filtering rules to ensure that the data aligns with the mandates of the respective authorities. A project conducted by ESMA in 2021 revealed that TRs seem to broadly follow the regulatory requirements concerning the provision of data to authorities as specified in EMIR mandates, which resulted in either underreporting or overreporting of EMIR data to authorities.[23]

3. Ways to boost data quality in EMIR

The introduction of data quality monitoring processes and use of automated tools have greatly increased the number of issues identified by the ECB and ESRB Secretariat and reported to ESMA via their data quality framework. The number of issues detected by other EU authorities is also rising sharply, thanks to constantly growing interest of the regulators in this dataset and development of analytical tools to analyse this rich information.[24]

This scaling up of data quality monitoring does not necessarily go in tandem with the number of issues solved. Supervisory actions often require laborious, and to significant extent manual process of contacting the reporting agent, explaining the issue, and agreeing on the way forward. Hence the supervisory channel has its limits, and it should be considered how to maximise our data quality "output", given limited resources. The following sections offer some ideas.

3.1 EMIR Refit – outlook on data quality

On 7 October the EU Commission published revised Regulatory and Implementing Technical Standards on EMIR reporting, under so-called EMIR Refit.[25] This move is part of a globally-

[22] See section 4.1.2.1 in European Securities and Market Authority (2023)

[23] See section 4.1.3.2 in European Securities and Market Authority (2023)

[24] See: European Securities and Markets Authority (2024)

[25] See:

- Commission Delegated Regulation (EU) 2022/1855
- Commission Delegated Regulation (EU) 2022/1860
- Commission Delegated Regulation (EU) 2022/1858

coordinated initiative of major jurisdictions to improve the transparency of the derivatives markets, and is aligned with international harmonisation efforts under the auspices of the FSB, CPMI, and IOSCO to foster global harmonisation of OTC derivatives reporting.

The changes to EMIR went into effect on 29 April 2024 and constitute a major overhaul of the reporting framework. They included, among others:

- Standardisation of the format of the reports submitted to TRs, in line with an ISO 20022 XML message, developed jointly by major global regulators;
- Alignment with international harmonisation efforts (CDE, UPI, UTI);[26]
- Additional granular information to be reported (203 fields vs. 129 currently);
- More detailed and robust event model;
- Mandatory, automated feedback from TRs to reporting agents;
- More robust validation and reconciliation procedures of the TRs, defined in a dedicated regulatory technical standard on data quality (EU Regulation 2022/1858);
- Comprehensive guidelines for reporting agents and TRs.[27]

The focus of the legislators on the data quality in EMIR Refit is very welcome and has the potential to materially improve the quality of the information reported in the medium term. The TRs will now have to implement procedures to ensure a certain level of quality of the information reported, and to provide respective feedback to the reporters on certain types of problems identified in the data. Harmonisation of the input messages submitted to TRs may present some challenges to the reporting entities at the beginning, but will lead to more standardisation in the EMIR reporting and reduce the risk of "translation" errors by the TRs.[28] Alignment with global reporting standards, in turn, will facilitate reporting of cross-jurisdictional entities.

While the steps taken by EMIR Refit will certainly enhance data quality in certain areas of reporting, we believe that there is still ample room for improvement. With relatively simple

---

[26] UPI = Unique Product Identifier

UTI = Unique Transaction Identifier

CDE = Critical Data Elements

For details, see: https://www.leiroc.org/international_bodies.htm

[27] See: https://www.esma.europa.eu/sites/default/files/2023-10/ESMA74-362-2281_Guidelines_EMIR_REFIT.pdf

[28] Before Refit, TRs implemented proprietary input formats, while they were obliged to transmit the data to the regulators using a standardized ISO20022 message. This led to issues, where information correctly reported by the entities was not adequately represented in the reports sent to the regulators.

measures the framework can be further enhanced to promote better reporting practices, simplify the reporting supervision, and ensure that the regulators receive data that can be effectively used for the exercise of their tasks.

3.2 Focus on what is important!

As described in section 2.3, data quality issues encountered in EMIR dataset that are systematic tend to follow specific patterns. Often, they apply to specific counterparties and product types and are recurring. A big part of an implausible aggregate can be linked to an individual counterparty, or even individual trade. In those cases, approaching the few largest misreporting entities and making them improve their reporting may bring considerable benefits, while keeping the effort limited.

A good example of this phenomenon is the evolution of the aggregate notional of CCPs vs. their clearing members during early 2024. At the beginning of the year, the aggregate notional reported by clearing members was around one third higher that the corresponding value reported by CCPs. This discrepancy presented a major challenge for the users of the data, putting into question the accuracy of the data on the notional of cleared trades. Due to only two counterparties improving their reporting, this difference shrank to only about 5% at the end of April. With this, the reliability of data on notional increased significantly.

3.3 Aggregated feedback to reporters

In many cases, reporting agents are not fully aware that the information that they show to regulators does not reflect the economic reality of their business. Often, an instant glance at the aggregated data derived from granular reporting reveals probable flaws in reporting. This may have different reasons: the granular information stored in companies' risk management systems may be incorrect, or the process to translate this internal information into reports (often designed and/or run by external providers) introduces errors that are propagated to the reports for regulators.

As described in section **Erro! A origem da referência não foi encontrada.**, under EMIR Refit the TRs are obliged to provide automated feedback to reporting agents, indicating trades where some specific information is missing, or numerical values are considered 'abnormal'. We see it a step that may improve the data quality, however, we also see certain flaws of this approach:

- the process to compile these reports is not standardized, but rather left to TRs to develop, which may lead to divergences in feedback provided by different TRs;

- the feedback will be very granular, hence there might be an incentive for the reporting entities to ignore it;
- information will be provided only for certain lifecycle events, not for outstanding trades – hence, once ignored, it will not appear again;
- such granular feedback will likely not catch certain systemic issues, that would be visible only on aggregate level.

Instead, or as a complement, a set of key aggregates computed by the TR from the information in the state report would be a solution for the shortcomings identified above. Such aggregates, similarly to what is reflected in the aggregates contained in the derivatives statistics compiled by the Bank for International Settlements,[29] would reflect basic and simple to compute quantities and breakdowns capturing key measures about the outstanding trades of individual entities. While one can think about different way to build such tables, our rudimentary proposal can be found in Tables 2 and 3 below.

Table 2: Proposed TRs' aggregated feedback to reporters – table A

Measure/breakdown	Asset class	Contract type	Notional currency[30]	Of which, with central counterparties
Count of trades				
Notional amount (EUR)				
Contract values (net)				
Contract values (gross)				

Table 3: Proposed TRs' aggregated feedback to reporters – table B

Measure/breakdown	Of which, with central counterparties
Initial margins posted	
Initial margins received	
Variation margins posted	
Variation margins received	
Excess collateral posted	
Excess collateral received	

Such tables would allow the reporting agents to quickly compare if the information being communicated to the regulators reflects the information they keep and use for internal purposes.

3.4 Standardization and detailed guidelines

3.4.1 TRs' compilation procedures

The history of EMIR reporting shows that providing detailed guidelines, validations rules, and making the counterparties reconcile between each other the information submitted to TRs can

[29] See https://data.bis.org/topics/OTC_DER

[30] A few main global currencies + "other" category

significantly improve the quality of data. EMIR Refit, with its detailed guidelines and procedures focused on data quality, is expected to be an important step in achieving better quality in the medium term.

We believe that this could be taken a step further. The EMIR legislation provides general guidance to TRs on how to compile some of their products, e.g.:

- Trade state and margin state reports
- Aggregated positions
- Automated feedback to reporters (so-called “warnings report”)

However, these guidelines often leave the detailed implementation to the TRs. This often leads to a situation where the information provided by different TRs is not fully aligned, limiting the usefulness of data and requiring manual and laborious cleaning of the data. Furthermore, each TR has to interpret the legislation separately, leading to unnecessary duplication of effort. Finally, this approach turns out to be error-prone, as the TRs may take assumptions, which are not consistent with the expectations of the regulators.

We are of the view that the guidance on compiling the TRs products should be provided in algorithmic, machine-readable format, in a way that removes any doubt on how the TRs’ output should be compiled from the information received from reporting entities. Such metadata-driven approach would minimize the risk of errors in TRs’ processing, and reduce the time and effort needed by TRs to introduce changes to their processing pipelines. This approach is even more justified now after the introduction of EMIR Refit, given that the input and output format of data processed by the TRs follow the same ISO 20022 standard.

These machine-readable guidelines should be public, allowing the TRs and regulators to scrutinize the procedures applied and better understand the data they receive and use for their daily tasks. This would also provide transparency to reporting agents, making them aware what happens to their data once they are submitted to TRs.

3.4.2 Reporting of valuations and margins

As explained in section 2.3, interpretation of data reported on contract values, margins and collateral presents considerable challenge to the regulators. The complexity of margining and valuation practices across different counterparties and limited standardisation of reporting makes it very difficult to map and quantify how risks and collateral are distributed and flow across entities in the financial system.

We believe that the two interlinked issues of valuation and collateral require a deeper reflection, among the regulators and market participants. A dedicated public consultation devoted to this

topic may provide a useful direction, making use of a more elaborate feedback from the industry and the regulators.

As a starting point for this discussion, we offer three general principles that, in our view, should be followed with the aim of achieving a clear and comprehensive reporting of valuation and margins:

- Instead of one field for 'collateral portfolio code', two fields – one for initial margin and another one for variation margin – should be introduced.
- Margin portfolios should be reported consistently across any pair of counterparties (including CCPs and clearing members). The portfolio codes should take the same values across the two counterparties to a trade. The fields linked to margins should be subject to intra-/inter-TR reconciliation of reported transactions.
- The valuation and margins should be reported as stock, not as a daily change.

4. Conclusions

EMIR is a granular, financial dataset of unprecedented richness, available to EU authorities. The quality of data remains very challenging, and considerable effort is necessary to use this information. Despite these barriers, the regulatory community intensively uses the data, which proved to be a crucial source of transparency in times of crises and market stress.

We are convinced that the data quality deficiencies should not serve as a discouragement for the potential users, but rather source of reflection on how this situation could be improved in a systematic and permanent manner. It is crucial to recognise the potential of this data. We believe that, with better data quality, we could see a rapid uptake in the use of the data, as well as increase of its use in automated monitoring pipelines.

This paper does not aspire to solve the problem of data quality in large-scale financial datasets, as this is a complex and nuanced topic that requires collaboration of multiple stakeholders. It should be rather treated as an initial contribution to the on-going discussion on how granular data should be handled and processed by the financial industry and regulatory community in rapidly developing technological landscape. We strongly believe that, by applying good practices, careful design of processes and making use of the available technology, significant benefits can be obtained, not only by regulators, but also market participants.

References

Agostoni, G. & de Charsonville, L. & D'Errico, M. & Leonte, C. & Skrzypczynski, G. (2023). Anomaly intersection: disentangling data quality and financial stability developments in a scalable way. https://www.bis.org/ifc/publ/ifcb59_32_rh.pdf

Agostoni, G. & Ianiro, A. & Jukonis, A. & Lenoci, F. & Letizia, E. & Skrzypczynski, G. (2024). Using trade-level derivatives data for macroprudential analysis. https://isi-web.org/sites/default/files/2024-03/ottawa-2023_ips-1077-Grzegorz%20Skrzypczynski.pdf

European Securities and Markets Authority (2023). 2023 Report on Quality and Use of Transaction Data. https://www.esma.europa.eu/sites/default/files/2023-04/ESMA74-427-719_2022_Report_on_Quality_and_Use_of_Transaction_Data.pdf

European Securities and Markets Authority (2024). 2023 Report on Quality and Use of Data. https://www.esma.europa.eu/sites/default/files/2024-04/ESMA12-1209242288-852_2023_Report_on_Quality_and_Use_of_Data.pdf

European Systemic Risk Board (2022). ESRB's view regarding data quality issues and risks for financial stability. https://www.esrb.europa.eu/pub/pdf/other/esrb.letter220713_on_data_quality_issues~18eccb6993.en.pdf

European Systemic Risk Board (2024). ESRB response to the consultative report by the BCBS, CPMI and IOSCO. www.esrb.europa.eu/pub/pdf/other/esrb.letter240417_response_iosco_consultation~a5c98d897b.en.pdf?f8b9fcf1e4cba54aa39782dbdc16d9ba

Pérez-Duarte, S. & Skrzypczynski, G. (2018). Two is company, three's a crowd: Automated pairing and matching of two-sided reporting in EMIR derivatives' data. https://www.bis.org/ifc/publ/ifcb49_51.pdf