26th Meeting of the Wiesbaden Group on Business Registers
Call for Papers
Neuchâtel, 24 - 27 September 2018
Zornitsa Manolova, Anja Lechner,
Global Legal Entity Identification Foundation (GLEIF)
Session No. 7
" Quality and Coverage of Statistical Business Registers "

Connect the dots with high quality LEI data. Introducing the GLEIF data quality management program

Abstract

The Global Legal Entity Identifier Foundation (GLEIF) is a supra-national, not-for-profit organization tasked to support the implementation and use of the Legal Entity Identifier (LEI). The LEI is a 20-digit, alpha-numeric code based on the ISO 17442 standard developed by the International Organization for Standardization (ISO). It connects to key reference information that enables clear and unique identification of legal entities participating in financial transactions. Simply put, the publicly available LEI data pool can be regarded as a global directory, which greatly enhances transparency in the global marketplace.

Today "data is the new gold", but simply gathering large amounts of data will in itself not serve to increase transparency across the global marketplace. Instead, what's needed is a free online source that provides open, standardized and high quality legal entity reference data with the potential to capture any entity engaging in financial transactions globally. GLEIF makes available such an open source with the Global LEI Index. It contains historical and current LEI records including related reference data in one authoritative, central repository and is accessible to LEI data users free of charge.

In the Global LEI System, GLEIF is responsible for monitoring and ensuring the high quality of LEI data. In cooperation with our partners, we focus on optimizing the quality, reliability and usability of the LEI data. This empowers market participants to benefit from the wealth of information available within the LEI population. This article gives an overview of GLEIF's data quality management and how we measure data quality within the Global LEI Index.

Our data quality management goal is to provide trusted, open and reliable LEI and legal entity reference data. A data quality cycle is used to achieve the data quality management objectives. Each data quality cycle step is performed through assigned quality gates:

- *Plan – data discovery and profiling.*
- *Do – data quality rule setting.*
- *Check – data quality monitoring and reporting.*
- *Act – data remediation.*

In close dialog with the LEI Regulatory Oversight Committee and the LEI issuing organizations, GLEIF has defined a set of measurable quality criteria to clarify the concept of data quality relative to the LEI population. For this, we have used standards developed by ISO. Example criteria include the completeness, comprehensiveness and integrity of the LEI data records. By

instituting a set of defined quality criteria, we have established a transparent and objective benchmark to assess the level of data quality within the Global LEI System.

The data quality checks defined in the published rule setting are assigned to one specific data quality criterion and describes one of the three possible data maturity levels: required quality, expected quality and excellent quality. GLEIF has also developed a methodology to score the level of LEI data quality. Details on the methodology applied to measure data quality in the Global LEI System are the topic of this article. In addition, the author highlights the importance and engagement of the business registry community for further improvement of the data quality in the Global LEI index.

1. Motivation – Connect the dots with high quality LEI data

The Global Legal Entity Identifier Foundation (GLEIF) is a supra-national, not-for-profit organization tasked to support the implementation and use of the Legal Entity Identifier (LEI). The LEI is a 20-digit, alpha-numeric code based on the ISO 17442 standard developed by the International Organization for Standardization (ISO). It connects to key reference information that enables clear and unique identification of legal entities participating in financial transactions. Simply put, the publicly available LEI data pool can be regarded as a global directory, which greatly enhances transparency in the global marketplace.

In May of this year, GLEIF published a new research paper "A New Future for Legal Entity Identification"[1] about invested valuable time and resource on client identification in the Financial Services Sector. A key finding in our research is the simple fact that financial institutions are using an average of four identifiers to accurately identify and crosscheck new legal entities throughout the client relationship. Approximately one-third of the respondents revealed that they're actually using a combination of five or more identifiers. Furthermore, over half (54%) of the survey participants agreed that the use of different identifiers for the same legal entity leads to inconsistency of information.

Today "data is the new gold", but simply gathering large amounts of data will in itself not serve to increase transparency across the global marketplace. Instead, what's needed is a free online source that provides open, standardized and high quality legal entity reference data with the potential to capture any entity engaging in financial transactions globally. GLEIF makes such an open source with the Global LEI Index available. It contains historical and current LEI records including related reference data in one authoritative, central repository and is accessible to LEI data users free of charge. Ultimately, there should be one identity behind every business and having an LEI helps to achieve this objective.

[1] "LEI in KYC: A New Future for Legal Entity Identification": https://www.gleif.org/en/lei-solutions/lei-in-kyc-a-new-future-for-legal-entity-identification

2. Ensuring high quality data: Introduction to GLEIS structure and responsibilities

In the Global LEI System (GLEIS) there are three main participants in the LEI issuing and maintaining process – the legal entity, the LEI Issuer also known as Local Operating Unit (LOU) and GLEIF (see Figure 1). To ensure a high quality level of data, close interaction between the involved parties and clear definition of their responsibilities is required.

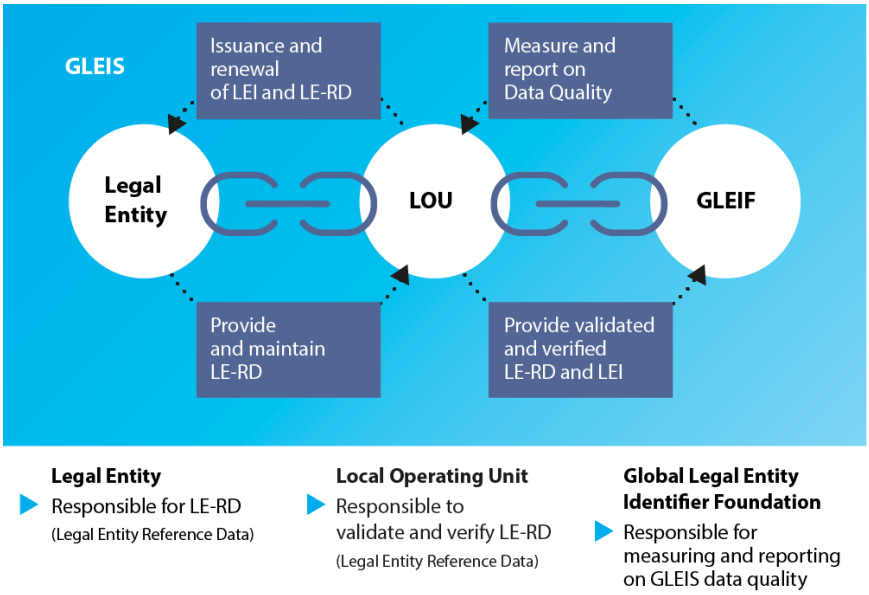

Figure 1: GLEIS Participants - Data Quality Value Chain

The process of ensuring LEI data quality starts with the registering entity (see Figure 2). Through self-registration, the content of the Legal Entity Identifier (LEI) data record is referred to as the legal entity reference data, and includes information such as addresses, legal jurisdiction etc. The legal entities (LEI owners) are responsible for providing accurate legal entity reference data and for making the LEI issuing organization aware of updates to the legal entity reference data.

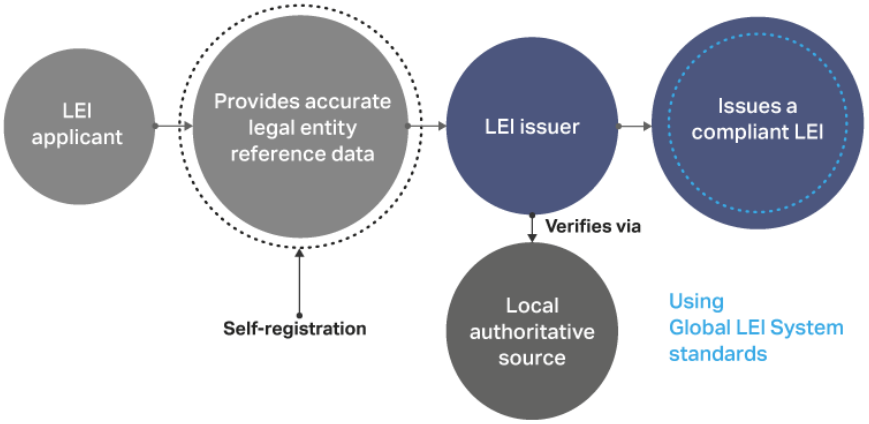

Figure 2: LEI Issuing Process

It is then the responsibility of the LEI issuing organization to validate and verify the legal entity reference data. In GLEIS, there is a clear distinction between the terms validation and verification. Validation is a safeguarding process to ensure that the inserted data satisfies the defined formats and additional input criteria, while verification means the provided data has been reviewed against local authoritative sources, e.g. a business register. This ensures LEI compliance with the LEI standards.

Data quality is also ensured via the annual LEI renewal process (see Figure 3). While the legal entity is required to notify the managing LEI issuing organization when changes occur to its legal entity reference data, the annual renewal process ensures that, at a minimum, the legal entity and the LEI issuing organization review and re-validate the legal entity reference data once a year.

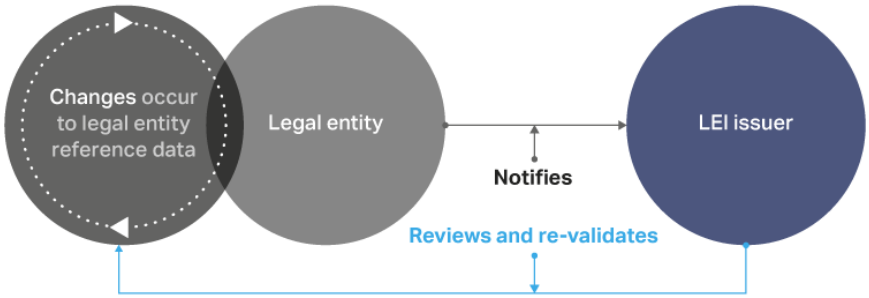

Figure 3: Annual Renewal Process

In the GLEIS, GLEIF is responsible for monitoring and ensuring the high quality of LEI data. In cooperation with its partners, GLEIF focuses on optimizing the quality, reliability and usability of the LEI data. This empowers market participants to benefit from the wealth of information available within the LEI population. The following chapter gives an overview of GLEIF's data quality management and how data quality is measured within the Global LEI Index.

3. GLEIF Data Quality Management program – data quality cycle and quality gates

GLEIF's Data Quality Management program is based on the idea of Total Quality Management, which was established by the Japanese manufacturing industry in the 1970s, to ensure the provision of high quality goods and services to customers. Since then both the theory and practice of Total Quality Management have been disseminated throughout the world and across different industries. In later years, the Massachusetts Institute of Technology established a research group on the topic of Total Data Quality Management[2]. They defined high data quality as "data that is fit for use by data user"[3].

Based on Deming's cycle of continuous improvement (see Figure 4) each data quality cycle step is performed through assigned quality gates:

[2] The MIT Total Data Quality Management Program. http://web.mit.edu/tdqm/www/about.shtml
[3] Strong, D., Lee, Y., and Wang, R. (1997, May). Data Quality in Context. Communications of the ACM, 40(5), pp. 103-110. http://mitiq.mit.edu/Documents/Publications/TDQMpub/StrongLeeWangCACMMay97.pdf

- Plan – data discovery and profiling.
- Do – data quality rule setting.
- Check – data quality monitoring and reporting.
- Act – data remediation.

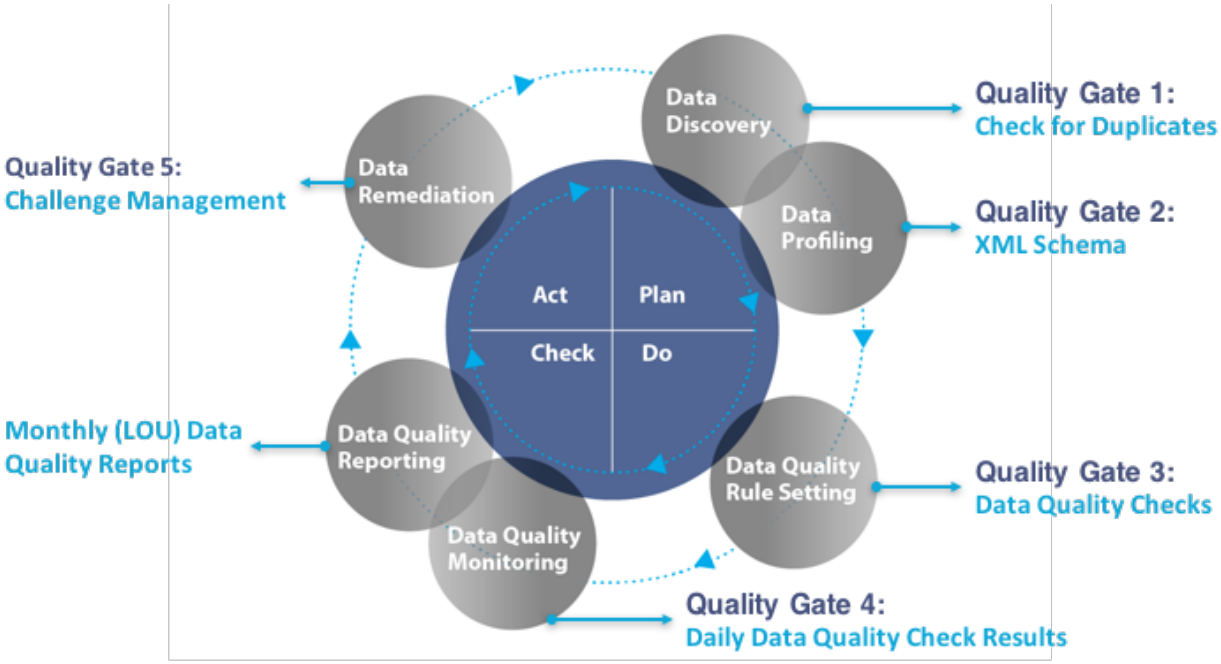

Figure 4: Data Quality Cycle

The first Quality Gate is represented by the "Check for Duplicates" facility to prevent the issuance of duplicates. GLEIF has developed a technical interface to support the identification of potential duplicates. The proven technology is based on a combination of three different state-of-the-art fuzzy-logic algorithms – Levenshtein, Monge-Elkan in combination with Cosine similarity and Cosine, used for name matching, surrounded by a series of pre- and post-processing steps. The applied approach focuses on precision of the results in order to minimize the number of false-positives and manual efforts for additional review. The facility is able to recognize potential uniqueness and exclusivity violations. The uniqueness of an LEI ensures that the same LEI is not issued twice for different entities, while the exclusivity of the record makes sure that one organization only obtains one LEI. Those characteristics of an LEI (to be unique and exclusive) align with the objective of having one identity behind every business.

The second Quality Gate includes the XML Schema. Its purpose is to prevent non-compliance to defined formats. Each reporting format is defined in a detailed specification document and XML schema definition (XSD) which enforces a minimum data quality (e.g. no spaces before first or after last word, correct enum values). A file which does not pass XSD validation cannot be included in the GLEIF Concatenated Files and the Global LEI Index. The schema defines:

- The structure of each data element.
- The associated code lists.
- The associated data element attributes.

The third Quality Gate is represented by the data quality checks and the corresponding Rule Setting. In close dialog with the LEI Regulatory Oversight Committee (ROC) and the LEI issuing organizations, GLEIF has defined a set of measurable quality criteria to clarify the concept of data quality relative to the LEI population. For this, standards developed by the International Organization for Standardization have been used. By instituting a set of defined quality criteria, GLEIF has established a transparent and objective benchmark to assess the level of data quality within the Global LEI System. The full list of 12 defined data quality criteria are displayed in Table 1.

Quality Criteria	Definition
Accuracy	The extent to which the data is free of identifiable errors / the degree of conformity of a data element or a data set to an authoritative source that is deemed to be correct or the degree the data correctly represents the truth about a real-world object
Accessibility	The extent to which data items that are easily obtainable and legal to access with strong protections and controls built into the process
Completeness	The degree to which all required occurrences of data are populated
Comprehensiveness	All required data items are included—ensures that the entire scope of the data is collected with intentional limitations documented
Consistency	The degree to which a unique piece of data holds the same value across multiple data sets
Currency	The extent to which data is up-to-date; a datum value is up-to-date if it is current for a specific point in time, and it is outdated if it was current at a preceding time but incorrect at a later time
Integrity	The degree of conformity to defined data relationship rules (e.g., primary/foreign key referential integrity)
Provenance	History or pedigree of a property value
Representation	The characteristic of Data Quality that addresses the format, pattern, legibility, and usefulness of data for its intended use
Timeliness	The degree to which data is available when it is required / concept of data quality that involves whether the data is up-to-date and available within a useful time frame; timeliness is determined by manner and context in which the data is being used
Uniqueness	The extent to which all distinct values of a data element appear only once
Validity	The measure of how a data value conforms to its domain value set (i.e., a set of allowable values or range of values)

Table 1: Data Quality Criteria

To measure the data quality criteria, checks have been defined based on the Common Data File (CDF) format. The full list of all data quality checks and the historical development of the Rule

Setting can be obtained from the GLEIF website[4]. These LEI checks are measured at different LEI data hierarchy levels (see Figure 5).

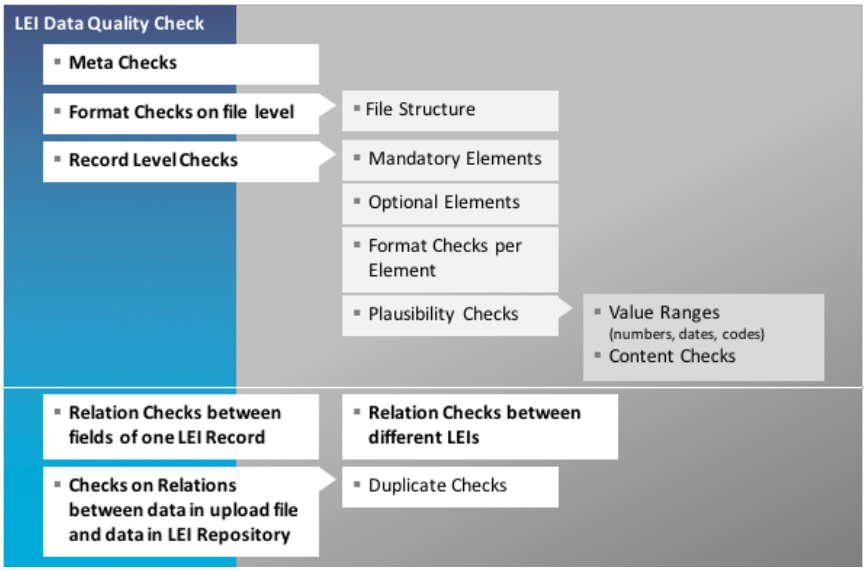

Figure 5: Data Quality Check Categories

Meta checks are not measured in the data file itself. These checks focus on timeliness, currency and accessibility of the data. The harder it is for the general public to access the information, the lower the accessibility. The more up-to-date the files that contain the relevant information are, the more current it is. The easier it is to access the information in a timely manner if it is available, regardless of timeframe, the more 'timeliness' it has.

Format checks are implemented on the file level, i.e. whether the files are compliant with the XML standard and Common Data File format. If a file is non-compliant to the standard, the information cannot be aggregated and therefore the data quality cannot be assessed.

Checks on the record level are applied to mandatory and optional field elements and cover format and plausibility checks (e.g. value ranges).

Additionally, there are several types of checks related to the different categories of relationships, which also need a different treatment:

- Relation checks between different fields of one record ensure business logic in the system.
- Relation checks between data in the upload file and data in the LEI repository: a prominent example of this type of relation check is the check for duplicates. These checks ensure internal consistency in the ecosystem and serve as a second level threshold of trust.

[4] https://www.gleif.org/en/lei-data/gleif-data-quality-management/about-the-data-quality-reports/supporting-documents#

- Relation checks between different LEIs are one of the most challenging checks, because they ensure compliance with business rules and require a huge cooperation effort from all involved parties: legal entity, LEI Issuers and GLEIF.

Each check is of the type 'If X then Y', where 'X' is described as a 'check precondition' and 'Y' is the 'check description'. If a record, relationship or exception does not fall into the 'check precondition', this check is 'not applicable'. If it passes the precondition and goes into the description and the value does not fulfil 'Y', the check is a fail (i.e. returns the value of 0). Quality criteria scores (Q_S) are the percentages of 'successful' and 'not applicable' data quality checks in relation to the total number of data quality checks for the respective quality criterion. The general formula for scoring the data quality criteria is the following:

$$Q_S = \frac{\sum_{i=1}^{I} q_i}{I}$$

Where:

- Q_S is the quality score for the respective quality criterion.
- q_i is the i^{th}, check result for the respective quality criterion with:

$$q_i = \begin{cases} 1 \; if \; check \; is \; \text{success} or \; \text{not applicable} \\ 0 \; if \; check \; is \; "failed" \end{cases}$$

- I is the total number of data quality checks performed for the respective quality criterion.

The Total Data Quality Score of the data quality criteria takes the average of the individual quality criteria scores (as previously mentioned Q_S). This average is not weighted by data quality criteria, meaning that each data quality criteria contributes equally to the total data quality score. The LEI Total Data Quality score (TQ_S) is therefore:

$$TQ_S = \frac{\sum_{S=1}^{N} Q_S}{N}$$

Where:

- TQ_S is the total data quality score.
- s in the summation is an index representing individual quality criteria.
- Q_S is the quality score for each respective quality criterion.
- N is the number of quality criteria for which there are checks implemented.

An additional concept implemented in GLEIS is that of the so-called Maturity Levels (see Figure 6). Maturity levels define the evolution of improvements in processes associated with what is measured. Therefore, they are scored differently from data quality criteria: while the scoring rules apply in a similar way, higher maturity levels can only be scored if the previous maturity level is fully achieved.

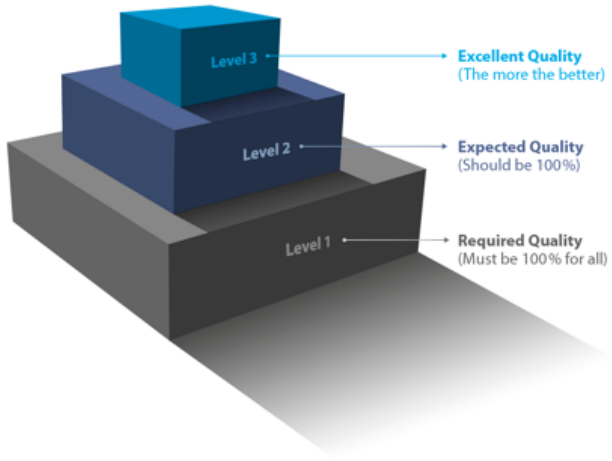

Figure 6: GLEIS Maturity Levels

At the moment, there are 75 checks active in Rule Setting and an additional 18 checks in implementation by the LOUs. Each check is assigned to exactly one criterion and exactly one maturity level. The mapping to the corresponding maturity level is done following the given rules below:

1 - Required: This level reflects repeatable success and is achieved when the following data quality checks are attained:

- all format checks on file level succeed.
- all record level checks regarding mandatory elements and format checks per element succeed.

2 - Expected: This level shows the managed success and is reached when the following data quality checks are passed:

- all record level checks regarding optional elements and plausibility checks succeed.
- all checks on relations between data in upload file and data in LEI repository succeed.

3 - Excellent: The third level is that of optimized success.

The fourth Data Quality Gate provides daily and monthly reports. The data quality checks are applied daily to each LEI Issuer's supplied data in common data file format (CDF). A data quality report is sent daily to each LEI Issuer detailing the results to ensure the defined quality criteria are achieved and to give feedback to the LEI Issuers about their performance on a daily basis. The monthly data quality reports communicate the overall quality in the LEI Index to the public, summarizing the results of GLEIF's assessments based on the above described criteria. The following two different types of monthly reports are produced and published on GLEIFs website:

- Global LEI Data Quality Reports: These reports demonstrate the overall level of data quality achieved in the Global LEI System.

- LEI Issuer Data Quality Reports: These reports analyze the level of data quality achieved by the individual LEI issuing organizations.

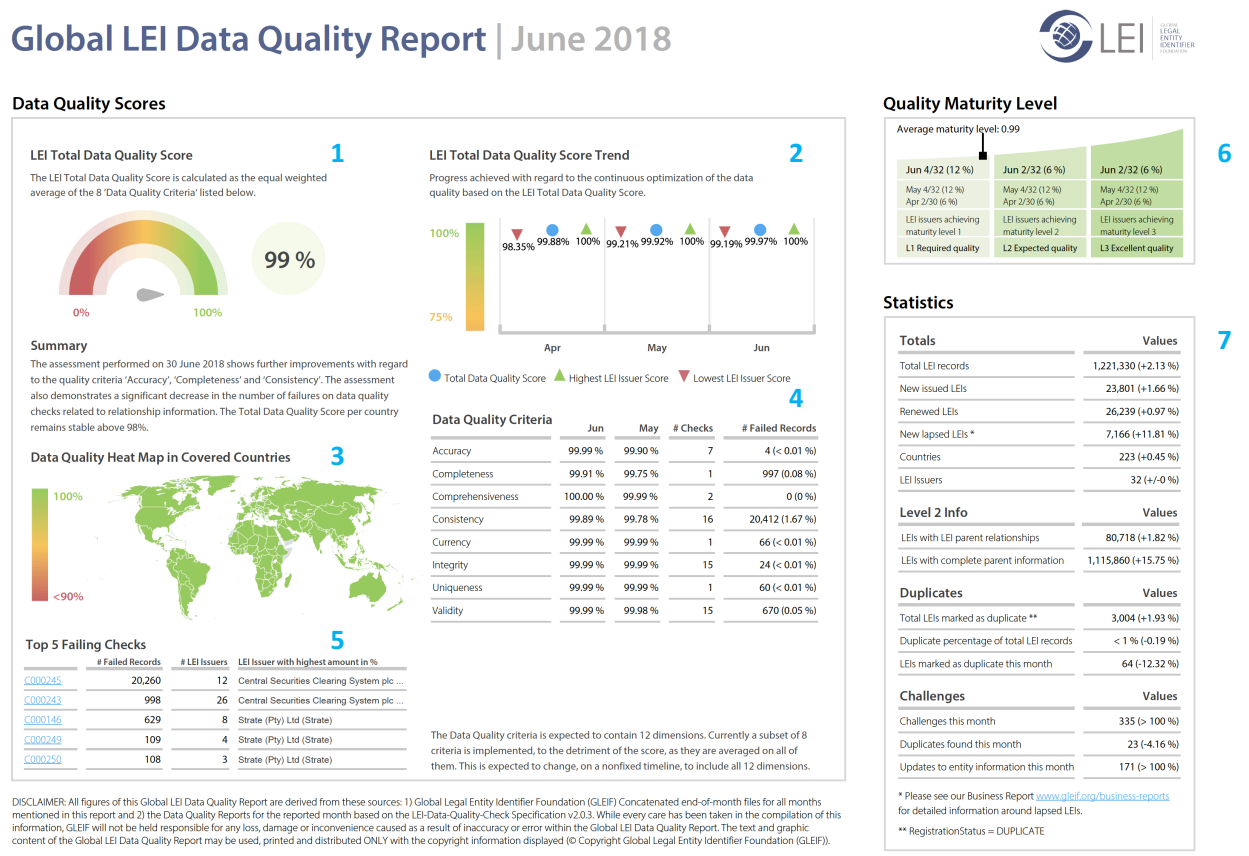

Figure 7: Global LEI Data Quality Report for June 2018

The Global LEI Data Quality Report (see Figure 7) includes seven main elements which provide the following information:

- The LEI Total Data Quality Score for the reporting period. (1)
- Progress achieved with regard to the continuous optimization of the data quality within the Global LEI System based on the LEI Total Data Quality Score. (2)
- The Total Data Quality Score per country achieved in the reporting period. (3)
- Results of GLEIF checks of the LEI data records against implemented quality criteria, i.e. the percentage of records that successfully passed the tests. (4)
- The section 'Top 5 Failing Checks' identifies those data quality checks performed by GLEIF which trigger the highest number of LEI records that fail these checks. In the report, the type of data quality check is indicated with a number. Please refer to the document 'Data Quality Rule Setting', available with the supporting documents, to learn which specific check corresponds to the number indicated with the report. (5)
- The percentage of LEI data records, which meet the requirements of distinct quality maturity levels. (6)

- Information on ‘Level 2’ data, duplicates and challenges for the reporting period. For background information: In May 2017, the process of enhancing the LEI data pool, by including Level 2 data to answer the question of ‘who owns whom’, began. This data allows the identification of the direct and ultimate parents of a legal entity and, vice versa, in order that the entities owned by individual companies can be researched. The ‘duplicates’ section in the report identifies the following issue: In line with applicable policy, one legal entity must only have one LEI. If it is identified that one legal entity has, for example, three LEIs, then two of these will be marked as duplicates. Duplicate LEIs are flagged in the Global LEI Index with the registration status. The centralized challenge facility made available by GLEIF extends the ability to trigger updates of LEI data to all interested parties. (7)

The last Data Quality Gate ensures continuous improvement by using crowd sourcing for Challenge Management (see Figure 8).

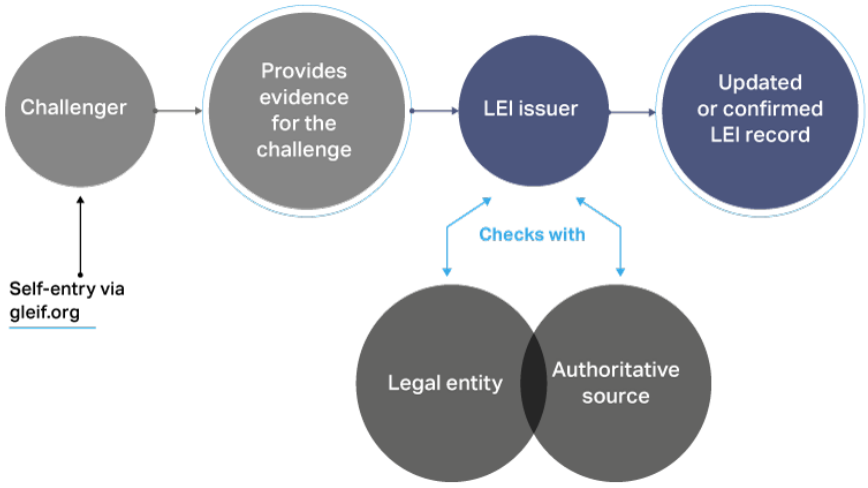

Figure 8: Challenge Process

The centralized challenge facility made available by GLEIF gives all interested parties the ability to trigger updates of an LEI’s data. Specifically, it offers an easy and convenient means to trigger the verification and, where required, speedy update of LEI records including related reference data. This centralized online service contributes further to ensuring that the publicly available LEI data pool remains a unique source of standardized information on legal entities worldwide. The GLEIF data challenge facility provides any user of LEI data with the opportunity to substantiate doubts regarding the uniqueness of an LEI code, the referential integrity between LEI records, or the accuracy and completeness of the related reference data. It also allows the indication of possible duplicate entries or any lack of timely response to LEI related corporate actions.

Once a challenge has been logged using the online form available on the GLEIF website, GLEIF immediately conveys the information to the relevant LEI issuing organization for follow-up. LEI

issuing organizations act as the primary interface for legal entities that have registered, or wish to obtain, an LEI.

At the time of this paper's publication, only one LEI data record can be challenged at a time, yet as many fields as needed within that record can be challenged. The process of challenging several LEI data records necessitates the entry of one challenge per record.

There is no complicated sign-in process necessary to submit a challenge. An email address must simply be provided in case contact is needed for further queries or in cases where evidence is required to underline the requested change. Should the correctness of an LEI data record be called into question through the GLEIF challenge facility, it is the responsibility of the relevant LEI issuing organization to resolve the matter in dialog with the impacted legal entity. If required, and subject to further verification against an authoritative source, the LEI issuer will update the information related to an LEI record. The aim is to resolve a challenge within ten business days.

To complement the established Quality Gates, GLEIF continuously monitors the data quality of the complete LEI data pool. Engaged users report to GLEIF about potential data quality issues as well. This enables GLEIF to maintain a list of topics that are addressed via data quality campaigns. Examples for such initiatives in the past were 'Unification of the syntax of city names' or 'Remediation of inconsistencies in the declaration of ultimate parentage'. Data quality issues are addressed via campaigns, when the majority of the LOUs are affected and a unified and collaborative approach is needed, since "Quality is everyone's responsibility" (W. Edwards Deming).

4. Conclusion – Importance and engagement of the business registry community

Quality cannot be adjusted overnight, neither can 100% quality be ensured. To achieve a high level of data quality, continuous and sustainable engagement is needed between all involved parties, alongside continuous communication with the data users.

Business registers are an important member of the LEI ecosystem, since they represent one of the main verification sources for LEIs and ensure the accuracy of the reference data. There are various scenarios in which GLEIF and the business registry community could engage and collaborate more closely to improve data quality further. One possible joint initiative could be ensuring the correct format of the registration authority entity IDs in the Global LEI Index. This would not only lead to an improvement in the quality of the LEI reference data but would also ensure the higher quality of future mapping exercises to additional identifiers - connecting the dots will enhance transparency in the global marketplace.

In addition, LEI Issuers combine the most up-to-date information from many authoritative sources, in order to ensure the most current and accurate information in the Global LEI Index. Thus, the LEI data pool represents an unified index of entity reference and relationship data across the world. This data could be leveraged by business registers themselves in situations, for example, where the business registers do not gather the full set of available information in the LEI index or the corporate actions information did not reach them.

The examples above present only two ways in which closer cooperation and collaboration can be achieved between GLEIF and the community of business registers to strengthen the Global LEI system. GLEIF is confident that many more relevant and productive collaboration scenarios could be explored.