Enabling global identity
Protecting digital trust

GLEIF Check for Duplicates

Dictionary



Public
Version 2.0 Final
2026-06-12

Table of Contents

1. Purpose of the Check for Duplicates.....5

2. Check for Duplicates workflow and methodology 6

2.1 Pre-processing..... 6

2.2 Filtering 7

2.3 Scoring..... 8

Version	2.0 Final
Date of version	2026-06-12
Created by	GLEIF
Approved by	Head of Data Quality Management and Data Science
Confidentiality level	Public

About this Document

This document contains a short description of GLEIF's Check for Duplicates facility providing more transparency to the users. This document does not replace the Implementation Note for Check for Duplicates or its Appendix.

Chapter 1 is about the purpose of the Check for Duplicates facility. Chapter 2 describes the current (as of the date of distribution of this document) approach for duplicates detection. The algorithm and parametrization may change in the future due to continuous improvement.

Change History

This section records the history of all changes to this document.

Date	Version	Description of change	Author
2026-06-12	2.0 Final	Final version after consultation	GLEIF
2026-03-31	2.0 Draft	Reflect Check for Duplicates 2.0 methodology	GLEIF
2021-11-10	1.2	Update layout and wording	GLEIF
2019-09-26	1.1	Clarifications	GLEIF
2016-10-06	1.0	Final version	GLEIF

Glossary

Term	Definition
Exclusivity violation	A violation of this type appears when any particular Legal Entity has sufficiently similar Reference Data against a predefined threshold.
LEI Database	The Global LEI Repository and other LEI Records that have been submitted to the Check for Duplicates facility.
LE-RD	Legal entity's reference data.
Uniqueness violation	A violation of this type appears when an LEI code appears more than once.
Vector embedding	In the context of the GLEIF Check for Duplicates, numerical representation of LE-RD computed on a selection of data elements that are deemed relevant for duplicate detection.
Weak tokens	Words found in entity names that are considered less important for distinguishing entities, compared to other parts of the legal entity's names. Typically, these weak tokens account for a significant number of characters and may therefore unnecessarily increase or decrease the similarity score in fuzzy matching algorithms.

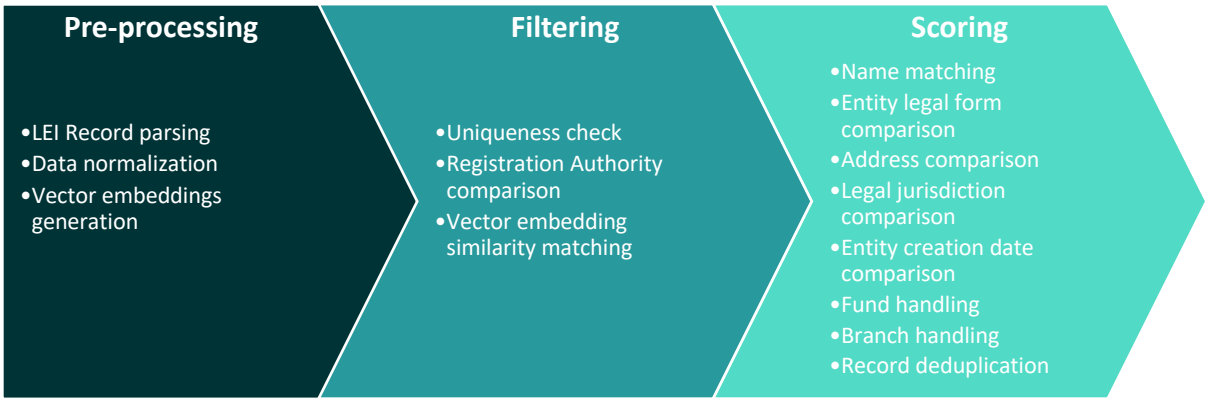
1. Purpose of the Check for Duplicates

In order to support the LEI issuing organizations, GLEIF provides the appropriate and mandatory processes and a technical interface for LEI issuers to check the LEI and LE-RD data for duplicate entries. LEIs are to be checked one at a time prior to their publication using an automated web service API (“Endpoint”) provided by GLEIF.

In addition to LEI issuers’ internal duplicate detection systems, LEI issuing organizations are required to use GLEIF’s Check for Duplicates facility to assess all LEI Records independent of their Registration Status prior to publication. To prevent duplicate LEI Records, all new records are compared during the issuance process against the full Global LEI Repository, as well as against not yet issued LEI Records submitted to the Check for Duplicates facility by other LEI issuers. This ensures that even if multiple LEI issuers are approached by the same legal entity, the LEI issuers will idenfity that multiple LEI Records are in the process of being issued for the same entity. In such cases, the LEI issuers coordinate with each other and with the client to prevent the introduction of duplicate records in the LEI system.

2. Check for Duplicates workflow and methodology

The general workflow of Check for Duplicates consists of three main steps:



The pre-processing contains a series of preparatory activities on the data elements prior to record comparison, such as removing punctuation and eliminating multiple consecutive spaces. This step also includes technicalities, such as parsing the LEI Record to extract the LE-RD and generating the vector embedding needed for the filtering and the subsequent duplicate detection.

The filtering step of the Check for Duplicates facility uses an AI-enhanced workflow to improve the duplicate detection. Vector embeddings are generated for each LEI Record in the LEI Database based on selected data elements relevant for duplicate detection. Similarly, LEI Records submitted to the Check for Duplicates facility are also converted into vector embeddings. Using similarity metrics, these embeddings are compared against existing vector embeddings to identify an initial set of suspected duplicates.

The initial set of suspected duplicates is pruned during the scoring step. This minimizes the amount false positives returned by the Check for Duplicates facility.

2.1 Pre-processing

The LEI Record is parsed to extract relevant data elements for duplicate detection (e.g., entity names, addresses, legal jurisdiction) used in the filtering and scoring stages. The data is then normalized through the following steps:

- Removal of multiple consecutive spaces, as well as leading and trailing spaces
- Removal of “Weak tokens”, such as legal forms and abbreviations
- Harmonization of character casing
- Removal of punctuation characters

Different stages of the filtering and scoring apply varying normalization criteria.

2.2 Filtering

Step 1: Uniqueness check

First, a uniqueness check based on the LEI code is performed and records with identical LEI codes to the submitted LEI Record are returned as Uniqueness violations in the response with "type": "UNIQUENESS". Regardless of the result of this check, the next steps are also performed for all records.

Step 2: Registration Authority comparison

An exact match on the RegistrationAuthorityID (RAID) and normalized RegistrationAuthorityEntityID (RAEID) will be suggested as a potential duplicate in the response with "type": EXCLUSIVITY. Likewise, an exact match on the (Other)ValidationAuthorityID ((O)VAID) and normalized (Other)ValidationAuthorityEntityID ((O)VAEID) will also result in a potential duplicate. Reserved codes for RegistrationAuthorityID and ValidationAuthorityID (RA888888 and RA999999) are excluded as well as missing RAEID and VAEID codes. In case of potential duplicate matches for FUND or BRANCH entities, additional steps are performed to reduce the number of false positives.

Example 1: Registration Authority comparison

LEI sent to the facility	LEIs in the Database	Potential Duplicates
RAID: RA123456 RAEID: AB-123.456.789	LEI A: RAID: RA123456 RAEID: AB123456789 LEI B: RAID: RA654321 RAEID: AB-123.456.789	**LEI A**

Note: LEI issuers should avoid publishing "n/a" (or similar) as RAEID for registries that do not provide an entity ID as this would count as exact match and potentially return a large number of suspected duplicate records. The usage of placeholder values like "n/a" is monitored via GLEIF's Data Quality Checks.

Step 3: Vector embedding similarity matching

The LEI Record submitted to the Check for Duplicates facility is converted into a vector embedding representation. Only a selection of data elements relevant for duplicate detection, such as entity names and addresses, are considered. Using similarity metrics, the LEI Record's vector embedding representation is matched against existing vector embeddings of LEI records in the Global LEI Repository and against LEI records that are in the process of being issued. The LEI Records most similar in their vector embeddings to the submitted record are further processed as part of the post-processing step.

2.3 Scoring

The scoring is performed on all records found in the filtering step (Registration Authority comparison and vector embedding similarity matching) in order to remove false positives from the list of potential duplicates returned to LEI issuers. The scoring process consists of several checks, each focusing on specific data elements of the LE-RD and contributing to the overall similarity score by either raising or lowering it. At the end of the process, all records with a similarity score below the predefined threshold are removed from the list of suspected duplicates. The remaining records are returned as Exclusivity violation in the response, with "type": "EXCLUSIVITY".

The main scoring steps are summarized below.

1. **Name matching**: In this step, the pre-processed names from the data elements LegalName, OtherEntityName and TransliteratedOtherEntityName are compared against those of each suspected duplicate, in all possible combinations. The comparison leverages a variety of string similarity algorithms to capture both character-level and token-level similarities:

    - Various normalized Levenshtein similarity variants
    - Normalized Jaro-Winkler similarity
    - Normalized similarity of leading character sequences (prefixes) of entity name

    Combining these algorithms allows for tackling a wide range of cases and ensures robustness. For example, high similarity may be detected if:

    a. Any of the names fully match those of the submitted record or are highly similar after whitespace and punctuation removal
    b. Any of the names match or are highly similar after removing "weak tokens" (e.g., legal forms)
    c. Any of the names match or are highly similar in their token composition

    The similarity score resulting from the name matching serves as the baseline for the subsequent filtering checks. Depending on their outcomes, the similarity score will be adjusted upward or downward in a sequential manner.

Example 2: Full name match irrespective of capitalization and punctuation

LEI sent to the facility	LEIs in the Database	High similarity
LegalName: YES Bank Inc.	LEI A: LegalName: yes-bank, Inc LEI B: LegalName: The Bank LTD	**LEI A**

In the context of the GLEIF Check for Duplicates, weak tokens may be removed at various stages of the scoring process. These tokens typically include legal form terms, abbreviations as well as other frequently used words, such as those found in fund entity names.

Example 3: Fuzzy name match including removal of weak tokens

LEI sent to the facility	LEIs in the Database	High similarity
LegalName: Flux Project, LLC LegalJurisdiction: GB	LEI A: LegalName: Flux Project Limited Liability Company LegalJurisdiction: GB LEI B: LegalName: Other Project, LLC LegalJurisdiction: GB LEI C: LegalName: Flux Project GmbH LegalJurisdiction: DE LEI D: Legal Name: Flex Project LegalJurisdiction: GB	**LEI A and LEI D**

In example 3, the weak tokens represent the legal forms and their abbreviations, allowing detection of suspected duplicates even if legal form are abbreviated. Weak tokens are handled per jurisdiction, therefore “GmbH” is not considered a weak token in the context of LegalJurisdiction GB.

Example 4: No false positive match thanks to removal of weak tokens

LEI sent to the facility	LEIs in the Database	High similarity
LegalName: 123 Handelsbolag	LEI A: LegalName: XY Handelsbolag LEI B: LegalName: Other Handelsbolag	**<none>**

Example 4 “Handelsbolag” which translates to “general partnership” is removed from the entity names during the scoring step.

2. **Entity Legal Form comparison**: If the pre-processed Entity Legal Form of the submitted LEI Record differs from that of a suspected duplicate, the baseline similarity score is reduced. This

comparison uses the Entity Legal Form (ELF) Code. If available, the OtherLegalForm element is also used. If no Entity Legal Form is available, this check is skipped.

3. **Address comparison**: This step uses the pre-processed addresses from the data elements LegalAddress, HeadquartersAddress, OtherAddress and TransliteratedOtherAddress in all possible combinations. High similarity increases the overall similarity score, while low similarity reduces it.
4. **Legal Jurisdiction comparison**: The submitted LEI Record and each suspected duplicate are compared to determine whether they share the same LegalJurisdiction. If the LegalJurisdictions differ, the similarity score is reduced. If subdivisions are present, they are considered in the comparison: If one record has Legal Jurisdiction "US-DE" and the other uses "US-NY", the jurisdictions are considered to be different. If a record does not provide any Legal Jurisdiction, this check will be skipped.

Example 5: Post-processing based on legal jurisdiction

LEI sent to the facility	LEIs in the Database	High similarity
LegalName: Power Steels Private Limited Legal Jurisdiction: US	LEI A: LegalName: Power Steels Private Limited Legal Jurisdiction: IN LEI B: LegalName: Power Steels Private Limited Legal Jurisdiction: US-NY	**LEI B**

5. **Extended fund handling**: If a suspected duplicate has EntityCategory "FUND", stricter Prefix-Suffix matching operations are performed on the entity names. This reduces the number of false positives for fund series where the only difference in their names is the numbering. If the submitted record itself is a "FUND", this step is extended to all suspected duplicates, regardless of their Entity Category.
   In cases of high numbers of suspected duplicates, a similar step is also performed for non-fund records.

Example 6: Post-processing for FUND

LEI sent to the facility	LEIs in the Database	Potential Duplicates
LegalName: Investment Fund 5 EntityCategory: FUND	LEI A: LegalName: Investment Fund 4 EntityCategory: FUND LEI B: LegalName: Investment Fund 3 EntityCategory: GENERAL LEI C: LegalName: Investment Fund 5 EntityCategory: FUND	**LEI C**

6. **Extendend branch handling**: If the submitted record has EntityCategory “BRANCH”, any difference in the legal address country will result in the removal of the affected suspected duplicates from scope. The same logic applies if the submitted record is not a “BRANCH”: Any suspected duplicate with EntityCategory “BRANCH” and a different legal address country is removed from scope.

Example 6: Post-processing for BRANCH

LEI sent to the facility	LEIs in the Database	High similarity
LegalName: XYZ LLC EntityCategory: BRANCH LegalAddress/Country: GB	LEI D: LegalName: XYZ LLC EntityCategory: BRANCH LegalAddress/Country: GB LEI E: LegalName: XYZ LLC EntityCategory: BRANCH LegalAddress/Country: AU	LEI D

7. **Entity creation date comparison**: This check verifies whether the EntityCreationDate matches between the submitted record and the suspected duplicates. If a match is confirmed, the similarity score is increased. Only the date component of the timestamp is considered. The check is skipped for records that don't report any EntityCreationDate.
8. **Record deduplication**: As potential duplicates may be identified across several steps, the same suspected duplicate record could match on different criteria in the filtering step (e.g., Registration Authority comparison and vector embedding similarity matching), potentially leading to the same suspected duplicate being returned multiple times. To reduce the effort for LEI issuers, deduplication is applied so that only one instance per duplicate type of each suspected duplicate is returned.