Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Eighth IFC Conference on *"Statistical implications of the new financial landscape"*

Basel, 8–9 September 2016

Unique identifiers in micro-data management – the Centralised Securities Database (CSDB) experience[1]

Asier Cornejo Pérez, Javier Huerga, Frank Mayerlen, Johannes Micheler,
European Central Bank

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

Unique identifiers in micro-data management – the Centralised Securities Database (CSDB) experience

Asier Cornejo Pérez, Javier Huerga, Frank Mayerlen, Johannes Micheler[1]

Abstract

The world of statistics is quickly expanding from pure macro-data compilation to micro-data management. This development brings amounts of data not seen before in official statistics. Without standardised unique identifiers such data will never deliver all its potential and data processing will quickly become unsustainable. At the same time reality shows that less standardised data sources need to be handled increasingly at least temporarily, e.g. during the start-up phase of a new data collection which starts from legacy data sets which have not been subject to standardisation efforts. As there is often no time to wait for full standardisation, it is necessary to process fully standardised and less standardised data in parallel. Modern statistical systems need therefore to produce robust and usable results with slightly imperfect and non-standardised input data and at the same time fully accurate data with fully accurate and standardised input data.

The Centralised Securities Database (CSDB) is a security-by-security database which contains reference, price and ratings data for more than six million active debt securities, equity shares and investment fund units issued worldwide. Drawing from the CSDB experience this note provides insights into the issues raised by the partial lack of unique identifiers and also shows how these issues have been addressed. Particular attention is given to the International Securities Identification Number (ISIN) and the Legal Entity Identifier (LEI). In the absence of accurate entity identification name matching may be used as a bridging technology where the Jaro-Winkler distance applied on the names of entities is presented together with stylised examples of its application.

Keywords: unique identifiers, International Securities Identification Number (ISIN), Legal Entity Identifier (LEI), euro area, debt securities, security-by-security databases, name matching algorithm, Jaro-Winkler distance.

JEL classification: C81 Methodology for Collecting, Estimating, and Organizing Microeconomic Data • Data Access

[1] Directorate General Statistics, European Central Bank, e-mail: asier.cornejo_perez@ecb.europa.eu, javier.huerga@ecb.europa.eu. frank.mayerlen@ecb.europa.eu, johannes.micheler@ecb.europa.eu. The views expressed in this paper are those of the authors and do not necessarily reflect those of the European Central Bank.

We would like to thank all colleagues in the European System of Central Banks who contribute to the development and operation of the CSDB.

Unique identifiers in micro-data management – the Centralised Securities Database (CSDB) experience 1

1. Introduction 3

2. A brief description of the CSDB 3

3. The CSDB experience with identifiers 5

   3.1. CSDB and the unique identification of securities 5

   3.2. CSDB and the identification of entities 6

4. Entities identification without unique identifiers – description of a possible 'name matching' alternative 7

   4.1. Process description and stylised examples 7

   4.2. Use of automatic bridging technologies – lessons learnt 11

5. Conclusion 11

6. References 12

1. Introduction

The world of statistics is quickly expanding from pure macro-data compilation to micro-data management. This development brings amounts of data in very large volumes, growing data diversity, increase in data velocity and frequency together with shorter data production cycles. It is increasingly recognised that large scale micro-data is a necessary ingredient for enhanced efficiency and decision taking at all levels. At the same time it becomes evident that shortcomings in the efficient and standardised processing of micro data could become one of the fastest growing operational risks in statistics.

Without standardised unique (surrogate) identifiers large scale micro data will never deliver all its potential and data processing will quickly become unsustainable. At the same time reality shows that less standardised data sources need to be handled increasingly at least temporarily, e.g. during the start-up phase of a new data collection which starts from legacy data sets which have not been subject to standardisation efforts. As there is often no time to wait for full standardisation, it is necessary to process fully standardised and less standardised data in parallel. Modern statistical systems need therefore to produce robust and usable results with slightly imperfect and non-standardised input data and at the same time fully accurate data with fully accurate and standardised input data.

The Centralised Securities Database (CSDB) is a security-by-security database which contains reference, price and ratings data for more than six million active debt securities, equity shares and investment fund units issued worldwide. Drawing from the CSDB experience this note provides insights into the issues raised by the partial lack of unique identifiers and also shows how these issues have been addressed. Particular attention is given to the International Securities Identification Number (ISIN) and the Legal Entity Identifier (LEI). In the absence of accurate entity identification name matching may be used as a bridging technology where the Jaro-Winkler distance applied on the names of entities is presented together with stylised examples of its application.

The paper is organised as follows. Section 2 provides a brief description of the CSDB and Section 3 illustrates the CSDB experience with identifiers for securities and issuers, with a focus on the ISIN and the LEI. Section 4 presents the situation where no unique identifier is available together with the possible alternative of using bridging technologies to integrate different sources in an automated way. Section 5 provides some conclusions and recommendations.

2. A brief description of the CSDB

Operational since 2009, the CSDB is a security-by-security (s-b-s) database set-up to hold complete, accurate, consistent and up-to-date information on all individual securities relevant for the statistical and increasingly non-statistical use by the European System of Central Banks (ESCB). The CSDB covers debt, equity and investment fund securities together with the respective price, issuer and rating information. As an example CSDB contains reference data on securities (e.g. outstanding amounts, issue and maturity dates, type of security, coupon and dividend information, statistical classifications, etc.), issuers (identifiers, name, country of

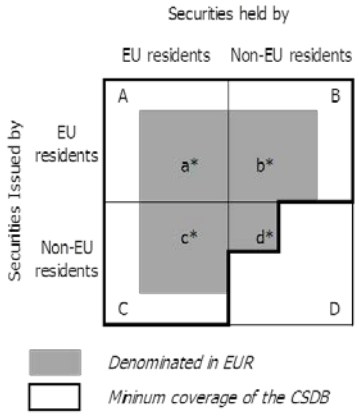
residence, economic sector, etc.) and prices (market valuation, estimated or defaulted). Moreover, the CSDB includes ratings information covering securities, securities issuance programmes, and all rated institutions (entities) independently of whether they are issuers of securities.

The CSDB covers securities issued by EU residents; securities likely to be held and transacted in by EU residents; and securities denominated in euro, regardless of the residency of the issuer and holders. Chart 1 illustrates this coverage. The CSDB currently contains information on over six million non-matured or "alive" debt securities, equities and mutual fund shares/units. In addition the database includes approximately nine million matured or "non-alive" securities (e.g. matured, early redeemed or cancelled). The number of issuers of "alive" securities is above 700,000.

Chart 1

Scope of securities covered in the CSDB

The total set falls within the heavy black line, and the sub-sets A to C. Coverage in D will be further extended as necessary, when transactions in or holdings of securities in this category are reported



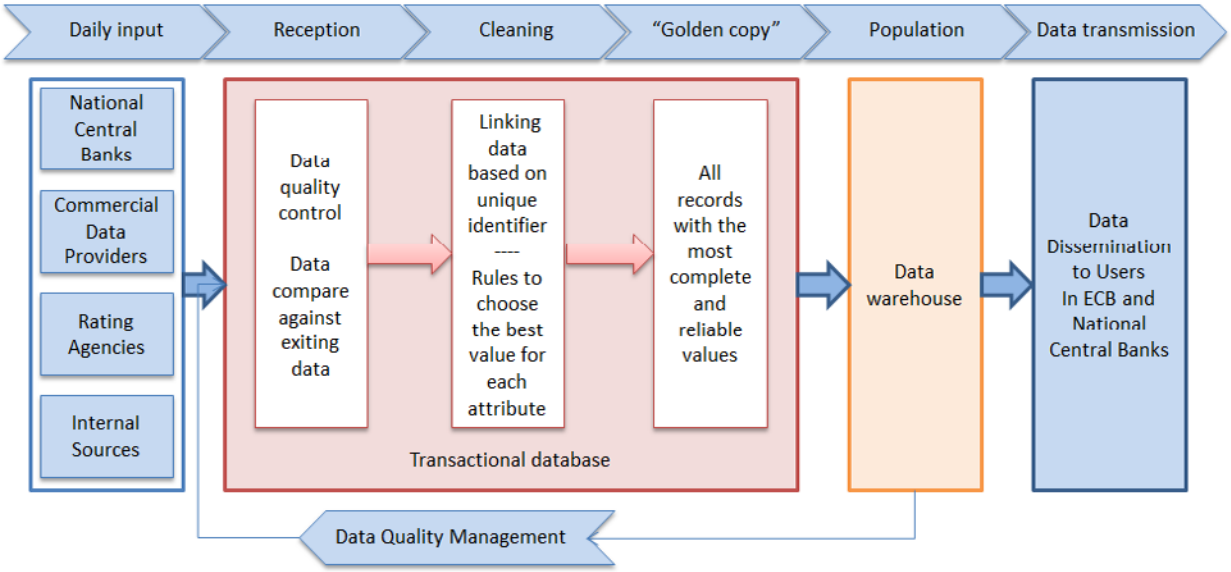
The CSDB is a multi-source system that receives approximately 2.5 million prices and 300,000 records on reference information per day from several commercial data providers. On a less frequent basis data are also provided by more than 20 National Central Banks (NCBs). Data is received automatically from all sources by the transactional system accessible via a web user interface, also called CSDB Portal. Upon reception, the data is compared against the existing or previously reported information. If unexpected changes are detected on the provision by one source, the respective data will be stopped to be further analysed by the system operators before being loaded. This avoids the loading of structurally faulty data and ensures data quality. Moreover, invalid data, i.e. data which are not compliant with the CSDB codelists, are filtered out at the start of the process.

Developed and hosted by the ECB, the CSDB is jointly operated by the members of the ESCB (the National Central Banks) and it is only accessible by them, i.e. there is no public access. Based on automatic algorithms the most reliable value for each attribute is selected by the system and gaps, in particular for prices and income, are filled with reliable estimates. To ensure the quality of CSDB data the system makes use of expertise within the ESCB in accordance with the respective Guideline of the

ECB on the data quality management framework for the CSDB (ECB/2012/21). The CSDB data processing including the data quality management loop is illustrated below.

Chart 2

Centralised Securities database information flow



Once data reception has finalised, the pooled data will usually contain some inconsistent information and so the data need to be cleaned. This "cleaning process" is automatically started by the system every evening. Cleaning is based on rules built into the system to choose the best (most reliable) value for each attribute in cases where sources might be contradictory. To finish the cleaning process, the data is enriched to fill in the gaps based on defined rules. After that CSDB data is made available in the data warehouse where a set of metrics has been created to detect inconsistencies and quality issues in the data. Different tools are made available to NCBs and the ECB to correct the data in a data quality management process that fixes the information at the root in the transactional database. This ensures data accuracy and consistency from the input to the output after finishing the data quality management process. Once the entire process is finished, data is made available to users at NCBs and at the ECB.

3. The CSDB experience with identifiers

3.1. CSDB and the unique identification of securities

The use of international standards in financial markets is not only necessary to ensure transparency but it is also one of the key features of s-b-s databases. While aggregate

data may include (and hide) different non-standardised components, the compilation of s-b-s data requires the existence and application of unique identifiers.

The CSDB uses the International Securities Identification Number (ISIN) as the unique identifier for grouping instruments information. This means that any information on securities must be reported together with the corresponding ISIN of the security. In this way it is ensured that any information referred to the same security is linked through the same ISIN. In the same vain, any security related output from the CSDB is identified by an ISIN which works as the 'primary key'.

As a multi-source system, the unique integration of input data in the CSDB is key for the quality of the output. The use of multiple identifiers for securities was previously tested in the CSDB and inevitable resulted in duplication of records referring to the same security. It was not always possible to find these duplicates as different identifiers have often only partial coverage of the total population. As an example provider 1 may send a security record identified only with the ISIN code and provider 2 sends a record referring to the same security identified only with a SEDOL code. Both records have no common identification and hence would wrongly exist as 2 distinct records in the system.

It is recognised that an approach relying only on the ISIN as a security identifier may omit from CSDB a small part of the scope of securities to be covered. However, it ensures that the most relevant securities are uniquely recorded in the CSDB with the proper data integration between the different sources. Fact-finding exercises carried out in the euro area reveal that debt securities without ISIN issued by euro area residents may amount to less than 1% of the total amount outstanding, although relative relevance may be higher in particular countries and sectors. The case of listed shares without ISIN in the euro area is irrelevant. Nevertheless the amount of other types of securities such as investment funds issued without ISIN may be somewhat more substantial although this is difficult to quantify due to the very lack of a unique international identifier. On balance, the risk of 'missing securities' in CSDB due to the non-availability of ISIN codes is considered much smaller than the risk caused by potential duplications.

3.2. CSDB and the identification of entities

The CSDB contains not only security data but also the link to the issuer of the security together with a number of attributes referring to the issuer. In the absence of a unique entity identification system for issuers the CSDB had to design and implement a procedure that permits to group securities belonging to the same issuer.

This procedure, so-called "grouping", is based on several relevant elements. In the first place, the input record information received by CSDB is at security level but this record also contains data on the issuer as available to the data provider, i.e. already providing the link between security and issuer. Second, the input record should contain an issuer identifier that is unique for each data provider, although different from provider to provider, as they are mainly proprietary codes[2]. Third, all possible

[2] Not all providers have a unique code to identify the issuer, therefore, several tools have been developed to overcome that additional restriction. First, when the code does not exist, the system creates, so called 'hash codes' based on the name and the country provided. Second, when the code

links between different issuer identifiers are created and can be indirectly matched through the ISINs i.e. different issuer identifiers linked to the same ISIN must correspond to the same security and hence to the same issuer. Finally, each issuer receives an unique internal CSDB identifier for system internal purposes together with all proprietary codes from each data provider.[3]

The creation of the Legal Entity Identifiers (LEI) offers new possibilities to substantially improve the entity identification process. In the first place the LEI could be used as additional code to perform the above described grouping process. On a more ambitious view the LEI should become the primary identifier for the grouping process while other identifiers would be redundant. Nevertheless even the universal use of the LEI will not solve all problems, given that data providers may still disagree on who is the issuer of a security. For that reason the CSDB grouping procedure is expected to stay in place for the foreseeable future, while noting that the optimal solution would be have to an unique authoritative source providing the correct link between ISIN and LEI.

4. Entities identification without unique identifiers – description of a possible 'name matching' alternative

As explained above the CSDB is based on unique identifiers, using the ISIN for securities and working towards fully using the LEI for entities. But, what can be done if no unique identifier is available for a data set? There is no doubt that establishing a compulsory unique identifier is the optimal solution, but what to do in the meantime? Combining several million records via non-standardised identifiers is considered a potential risk in terms of feasibility and workload. Against this background a machine-driven and automated data integration process without unique identifiers has been investigated.

This possible alternative relies on the comparison of non-unique and non-standardised identifiers such as e.g. names, country of residence, etc. The test implementation relied on existing CSDB processes and enhanced them with bridging technologies like the Jaro-Winkler distance. The concept relies on a centralised (and self-checking) automated data cleaning process which allows the integration of data coming from multiple sources according to a single unified procedure. The overall process should be highly scalable to handle potentially substantial amounts of data. The process used to proof the concept and the main conclusions are presented below.

4.1. Process description and stylised examples

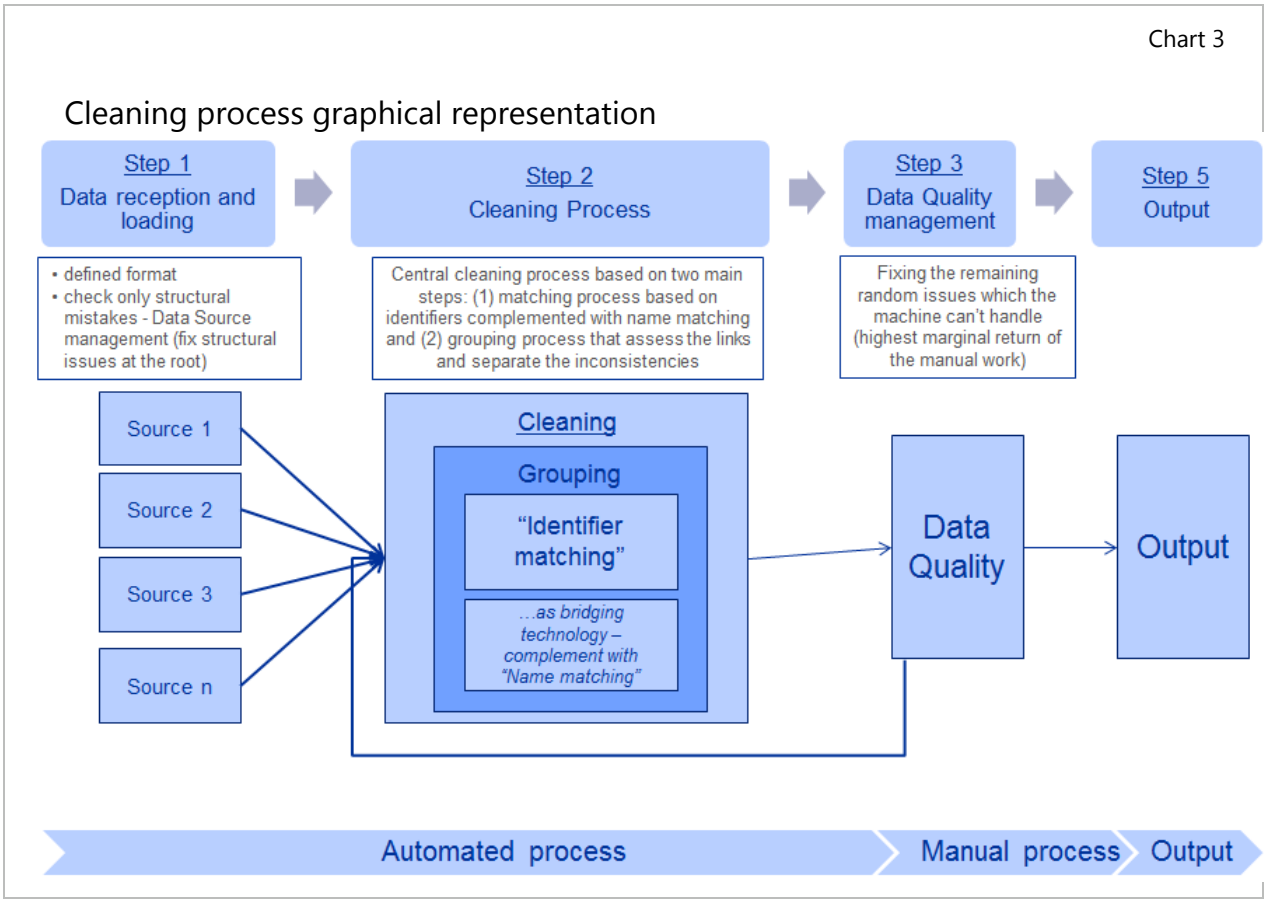
CSDB uses micro level information coming from different sources, which have identifiers that are often not unique and also not always shared across different data sources. The alternative process consists in two steps, first creating, e.g. based on the name, all feasible links between all records as provided by different sources and

is not unique for each issuer, the link between the instrument and the issuer is weaker but still used 'to join' securities issued by the same issuer if no other information is provided.

[3] In case of inconsistencies between different issuer identifiers the grouping procedure creates so-called "clash groups", i.e. cases in which different data providers disagree on who is the issuer of the security. These cases need to be resolved manually.

referring to the same entity. In a second step, a 'grouping' algorithm based on CSDB technology automatically verifies these links to create a unique representation of a single entity.

Chart 3



First step – creation of links between the data sources

For the first step and in the absence of common unique identifiers the CSDB uses a name matching algorithm to map ratings data from different ratings agencies. To significantly increase the total matching grade and complement the current lack of identifier matches, an improved name matching algorithm has been tested in a recent proof of concept.

The tested name matching algorithm is a string metric based on Jaro-Winkler. The Jaro-Winkler string metric takes into account the number of matching characters and transpositions of two strings and can take any value between 0 (no match) and 1 (perfect match). Other name matching algorithms that could have been used for the same purpose are, for example, the Levenshtein distance and the n-grams.

The name matching algorithm consists of three stages:

a) String unification: This step has two purposes. First, a context-free unification is applied based on an automated process that runs for all reported data (all string lower case, unification of special characters, unification of the legal form). Second, only the part of the entity name that is relevant to describe the company is highlighted. The idea is that non-relevant parts of the name should not dominate the name matching (for example for "BP Deutschland" vs. "BD Deutschland" the country name is longer than the actual name of the company and hence dominating).

b) Apply name matching algorithm and use some special selection mechanism based on the first string letters that have been proven to be significantly discriminating between true and false matches. A threshold can be implemented to ignore all counterparty pairs where the name matching grade of the Jaro-Winkler distance is less than a certain value, e.g. 0.9.

c) Final selection: Using reference data to make a final selection from all feasible pairs of the result from the first two steps.

With the above algorithm it becomes clear that the quality of the entity reference data from the different sources in terms of standardisation and coverage plays an important role in steering the 'bias/variance trade off' between high matching rate and quality. With high quality in the reference data, the threshold in step b can be reduced, allowing for a higher matching rate (with more false positives), knowing that in step c high quality reference data can be used in the final selection.

Second step – Entity Grouping: Bringing all records together which refer to the same entity

The term 'Entity Grouping' stands for the task of disambiguating instances of real world entities in various records by grouping. The identifier and name matching algorithms have already created a large set of links between the data provided by the different sources. The Entity Grouping clusters the links between them that correspond to the same entity. The grouping algorithm can be made subject to various constraints. In the case tested, the constraint is that one identifier can only refer to one entity. A violation of this constraint is an inconsistency that cannot be resolved automatically.

The CSDB has already implemented this entity grouping algorithm called 'Party Grouping' that automatically resolves multipartite links and creates clash free (free of constraint violations) 'Main Groups'. All links, where above constraint of the uniqueness of identifiers is violated are put into a 'Clash Group'.

Box 1

Stylised examples of data integration depending of the availability of identifiers

Case 0 – Instrument integration: Multiple sources of instruments with ISIN code

This first example shows the link of different sources based on the existence of a unique identifier, ISIN for the instrument that is used by all sources. Once the link is made based on the ISIN, the most complete information of the instrument can be obtained as a result of a compounding process.

	Output	Source 1	Source 2	Source 3
ISIN	XS1831830158	XS1831830158	XS1831830158	XS1831830158
ESA 2010 Classification	F.511	F.5	F.511	F.3
Issue date	05/09/2016	05/09/2016		05/09/2016
Nominal Currency	EUR	EUR	USD	EUR

Case 1 - Multiple sources of entities with LEI code

Similarly to the first example, Case 2 shows the link of different sources providing information of entities based on the existence of a unique identifier, LEI, used by all sources. Once the link is made, the compounding process allows having complete information of the entity.

	Output	Source 1	Source 2	Source 3
LEI	918184731989134130AB	918184731989134130AB	918184731989134130AB	918184731989134130AB
ESA 2010 Sector	S.12201	S.122	S.12201	S.122
Country	IE	IE		IE
Name	Bank One Limited	Bank One Limited	Bank One Ltd.	Bank One Limited

Case 2 - Multiple sources of entities with instruments with ISIN code but no LEI

When information of the entities is not provided with a common identifier, like LEI, other alternatives can be envisaged. The data provided for the instrument can be used as an indirect link to create the entity. This example shows how the availability of the ISINs allows to create a link between internal identifiers that are not of the same type. Once the link is created, the compounding process allows having complete information of the entity.

	Output	Source 1	Source 2	Source 3
ISIN	DE1831830143 AT3426754567 XS1831830143	XS1831830158 DE1831830143	DE1831830143 AT3426754567	XS1831830158
Identifier	X,Y,Z	X	Y	Z
Identifier type		Internal	Internal	Internal
ESA 2010 Sector	S.122	S.122	S.12	S.122
Country	IE	IE		IE
Name	Bank Two Limited	Bank Two Ltd.	Bank Two Limited	Bank Two Limited

Case 3 - Multiple sources of entities with no LEI - using name matching and common identifier types

If the LEI is not available and indirect links through instruments information are not possible, the proposed process described in section 4.1 could be used: link the information provided based on the same identifier types (source 1 and 2 link with the same VAT code) complementing it with a name matching algorithm that allows linking the source 3 based on a common name and other reference data like country. Once the link is created, the compounding process can be put in place.

	Output	Source 1	Source 2	Source 3
Identifier	D,Z	D	D	Z
Identifier type		VAT	VAT	Internal
ESA 2010 Sector	S.122	S.122	S.12	S.122
Country	IE	IE	IE	IE
Name	Bank Three Limited	Bank Three Ltd.	Bank three limited	Bank Three Limited
String matching		BankThree + IE	Bankthree + IE	BankThree + IE

4.2. Use of automatic bridging technologies – lessons learnt

There is evidence that the identification and integration of micro data information on entities involves large volumes of information. Therefore, the process described above has been tested as an enhanced concept of existing CSDB procedures. When dealing with large volumes of information, any process should rely on a fully automated and machine-driven approach. To guarantee this, it is required to rely on a unique, comprehensive and stable identification like the Legal Entity Identifier (LEI). Currently, the LEI is not yet fully used by all possible sources, therefore, non-standardised identifiers need to be used, provided that they are clearly defined and also consistently implemented, maintained and applied by the different sources.

The CSDB uses the ISINs as additional identifier to link the different sources, but the use of a name matching algorithm has also proven to be useful over the years. The use during the test of a more sophisticated name matching algorithm could mitigate to a large degree the risk caused by the current non-availability of LEI and could also demonstrate the feasibility of an automated and scalable solution. In that respect the tested process described in section 4.1 has shown that further efforts should be put on improving reference data of each entity which could be used in conjunction with the name matching to enhance its precision.

5. Conclusion

Unique identifiers play a key role in micro-data data bases. In the case of securities micro-data the ISIN and LEI are the most relevant identifiers. This paper has explained how and for what purpose the CSDB makes use of the ISIN and the LEI.

However, there are situations in which the unique identifiers do not (yet) exist, are not sufficiently established or have not yet sufficient coverage. In these cases it is necessary to use automatic procedures to overcome as much as possible the problem until sufficient coverage of the unique identifiers has been reached. A procedure to deal with the absence of any unique identifier in the case of entities is presented in this note, showing that automated data integration is possible. In doing so, the procedure presented does not only demonstrate the general feasibility of the concept but also shows how automation can be applied to the largest extent.

In more general terms, this note presents evidence that the identification and integration of micro-data on assets and entities at the required scale should rely as a backbone on a fully automated and machine driven approach. To guarantee this, it is also required to rely on a unique, comprehensive and stable identification like the ISIN and the LEI. In the short-term, sophisticated matching techniques, like the CSDB grouping or the name matching algorithms may as a bridging solution temporarily complement the identifier matching and help in mitigating the risk caused by the current non-full coverage of the LEI.

6. References

Cornejo Pérez, A, Huerga, J (2016): The Centralised Securities Database (CSDB) – Standardised micro data for financial stability purposes, IFC Bulleting No 41, Irving Fisher Committee, Basle, May.

ECB (2014): Opinion of the European Central Bank of 24 June 2014 on a proposal for a Regulation of the European Parliament and of the Council on reporting and transparency of securities financing transactions (CON/2014/49).

ECB (2015): Opinion of the European Central Bank of 17 March 2016 on a proposal for a regulation of the European Parliament and of the Council on the prospectus to be published when securities are offered to the public or admitted to trading (CON/2016/15).

ISO (2012): ISO 17442 (E) Financial services – Legal Entity Identifier (LEI).

ISO (2015): ISO 6166: 2013 (E) Securities and related financial instruments – International securities identification numbering system (ISIN).

European Central Bank (2010): The "Centralised Securities Database" in brief, Frankfurt am Main, February.

Winkler, W.E (1990): String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods, American Statistical Association, p. 354–359.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Eighth IFC Conference on *"Statistical implications of the new financial landscape"*

Basel, 8–9 September 2016

Unique identifiers in micro-data management – the Centralised Securities Database (CSDB) experience[1]

Johannes Micheler,
European Central Bank

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

EUROPEAN CENTRAL BANK

EUROSYSTEM

Johannes Micheler
External Statistics Division

Unique identifiers in micro-data management – the CSDB experience

IFC Biennial Basel Conference

8-9 September 2016

Disclaimer: The views expressed in the paper presentation are those of the authors and do not necessarily reflect those of the European Central Bank.

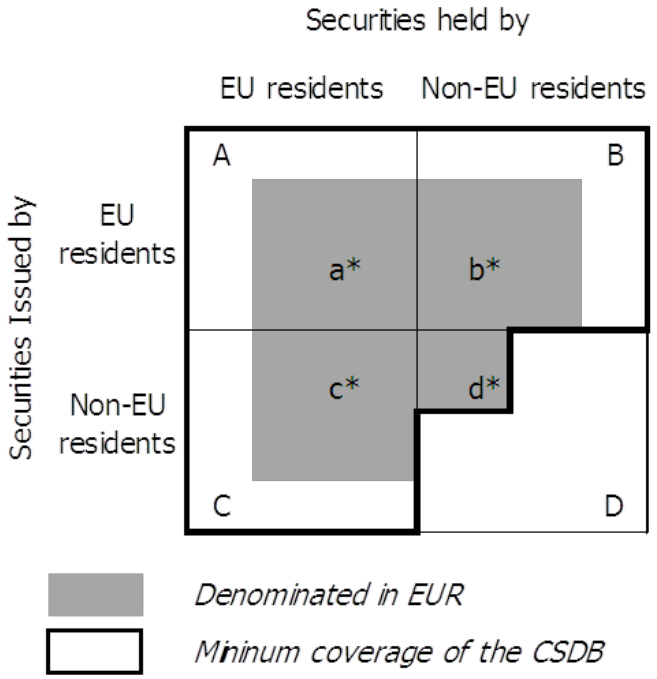
Overview

1. A brief description of the CSDB
2. Entity identification without unique identifiers
3. Conclusion

A brief description of the CSDB

The Centralised Securities Database (**CSDB)** is a securities reference database that holds **complete**, **accurate**, **consistent** and **up to date** information on all individual securities relevant for the statistical purposes of the ESCB.

- CSDB is **ISIN** based (International Securities Identification Number)
- Contains information on **more than 6 million alive debt** securities, **equities,** and **mutual funds** share/units
- Covers **instruments, issuers, corporate actions** and **ratings** data
- **Daily update** frequency
- Data received from **multiple sources** on same instrument and issuer are **grouped and compounded**
- CSDB located at the ECB and is shared by the ESCB



Entity identification without unique identifiers 1/6

CSDB and the unique identification of securities

- Compilation of **security by security data** requires a unique security identifier.
- The CSDB requires the **ISIN** to group and compound security reference data from multiple data sources.

CSDB and the identification of entities

- **No unique identifier with full global coverage exists yet for entities**
- An entity can currently be identified via
  - Proprietary identifiers from individual data sources
  - Standardised identifiers as the 'Legal Entity Identifier' (LEI)
  - CSDB artificial entity identifier (if no identifier is provided)

Entity identification without unique identifiers 2/6

CSDB party grouping
a multi partite entity resolution engine

- Multi partite entity resolution = **disambiguating** manifestations of entities provided by various data sources **by linking and grouping**.
  - **Step 1** - The CSDB uses the ISIN code to create all feasible **links between issuer identifiers**.
  - **Step 2** – The CSDB creates **a unique representation** (called 'Main Group') of the entity using the CSDB **party grouping**.

- If an inconsistent case occurs, the party grouping cuts the links that are causing the inconsistencies and creates a '**Clash Group**'.
- Identifiers of bad quality can be configured to not be able to create 'Clash Groups'.

Potential extension of the CSDB party grouping using name matching technologies

- Current party grouping algorithm requires **a 'glue'** (e.g. ISIN code) to create the set of **feasible links** between issuer identifiers.
- If the ISIN and LEI codes are not available → No links can be created → entities cannot be grouped.
- **Name matching algorithms** can help to **extend** the set of **feasible links** of **step 1**. Three stage approach is required.
  - **Stage 1: String unification**
  - **Stage 2:** Apply **name matching algorithm**, e.g. Jaro Winkler
  - **Stage 3: Final selection** based on entity reference data, e.g. country of residence, legal form, etc.
- Apply CSDB party grouping of step 2.
  - This step also helps identifying 'false positives', that cannot be dealt with by any name matching algorithm → see examples

Illustration of CSDB party grouping

- Case 1: Multiple sources of entities with LEI

	Main Group	Source 1	Source 2	Source 3
LEI	918184731989134130AB	918184731989134130AB	918184731989134130AB	918184731989134130AB
ESA 2010 Sector	S.12201	S.122	S.12201	S.122
Country	IE	IE		IE
Name	Bank One Limited	Bank One Limited	Bank One Ltd.	Bank One Limited

Illustration of CSDB party grouping

- Case 2: Multiple ISINs from 3 sources without inconsistencies

	Main Group	Source 1	Source 2	Source 3
related ISINs	XS1831830158 DE1831830143 AT3426754567	XS1831830158 DE1831830143	DE1831830143 AT3426754567	XS1831830158
Identifier	X,Y,Z	X	Y	Z
Identifier type		Internal	Internal	Internal
ESA 2010 Sector	S.122	S.122	S.12	S.122
Country	IE	IE		IE
Name	Bank Two Limited	Bank Two Ltd.	Bank Two Limited	Bank Two Limited

- Case 2a: Multiple ISINs from 2 sources with inconsistencies

	Main Group	Clash Group	Source 1	Source 2	Source 2
related ISINs	XS1831830158 DE1831830143	AT3426754567	DE1831830143 XS1831830158 AT3426754567	DE1831830143 XS1831830158	AT3426754567
Identifier	X,Y	X,Z	X	Y	Z
Identifier type			Internal	Internal	Internal
ESA 2010 Sector	S.122	S.12	S.122	S.122	S.12
Country	IE	AT	IE		AT
Name	Bank Two Limited	Bank Two Ltd.	Bank Two Ltd.	Bank Two Limited	Bank Two Limited

Illustration of CSDB party grouping including name matching

- Case 3: Multiple entities from three sources without inconsistencies

	Main Group	Source 1	Source 2	Source 3
Identifier	D,Z	D	D	Z
Identifier type		VAT	VAT	Internal
ESA 2010 Sector	S.122	S.122	S.122	S.12
Country	IE	IE		AT
Name	Bank Two Limited	Bank Two Ltd.	Bank Two Limited	Bank Two Limited
String matching		banktwo + IE + LTD	banktwo + IE + LTD	banktwo + IE + LTD

- Case 3a: Multiple entities from two sources with inconsistencies

	Main Group	Clash Group	Source 1	Source 1	Source 2
Identifier	B,X	A,X	A	B	X
Identifier type			Internal	Internal	Internal
ESA 2010 Sector	S.122	S.122	S.122	S.122	S.122
Country	IE	IE	IE	IE	
Name	Tranche I	Tranche 1	Tranche 1	Tranche 2	Tranche I
String matching			tranche1+IE	tranche2+IE	tranchei+IE

Conclusion

- **Unique identifiers** play a **key role** in **micro-data management**. In the case of s-b-s data, **ISIN** and **LEI** are most relevant. Both should ideally have full coverage, i.e. **become compulsory**.

- Technologies that allow to **compile usable data** using **incomplete and non-standardised data provision** are key to bridge the gap until fully standardised data is available.

- The **CSDB party grouping** based on the ISIN has proven to produce **reliable results** in terms of **coverage** and **data quality**.

- A **pilot exercise** in the year 2016 has proven that the CSDB party grouping technology could be upgraded using **name matching techniques** to further **increase the data coverage** while keeping the same degree of **quality**.