Benefits and Challenges of Data Linking:

U.S. experience linking data on foreign-owned U.S. companies to domestic employment data

Patricia Abaroa

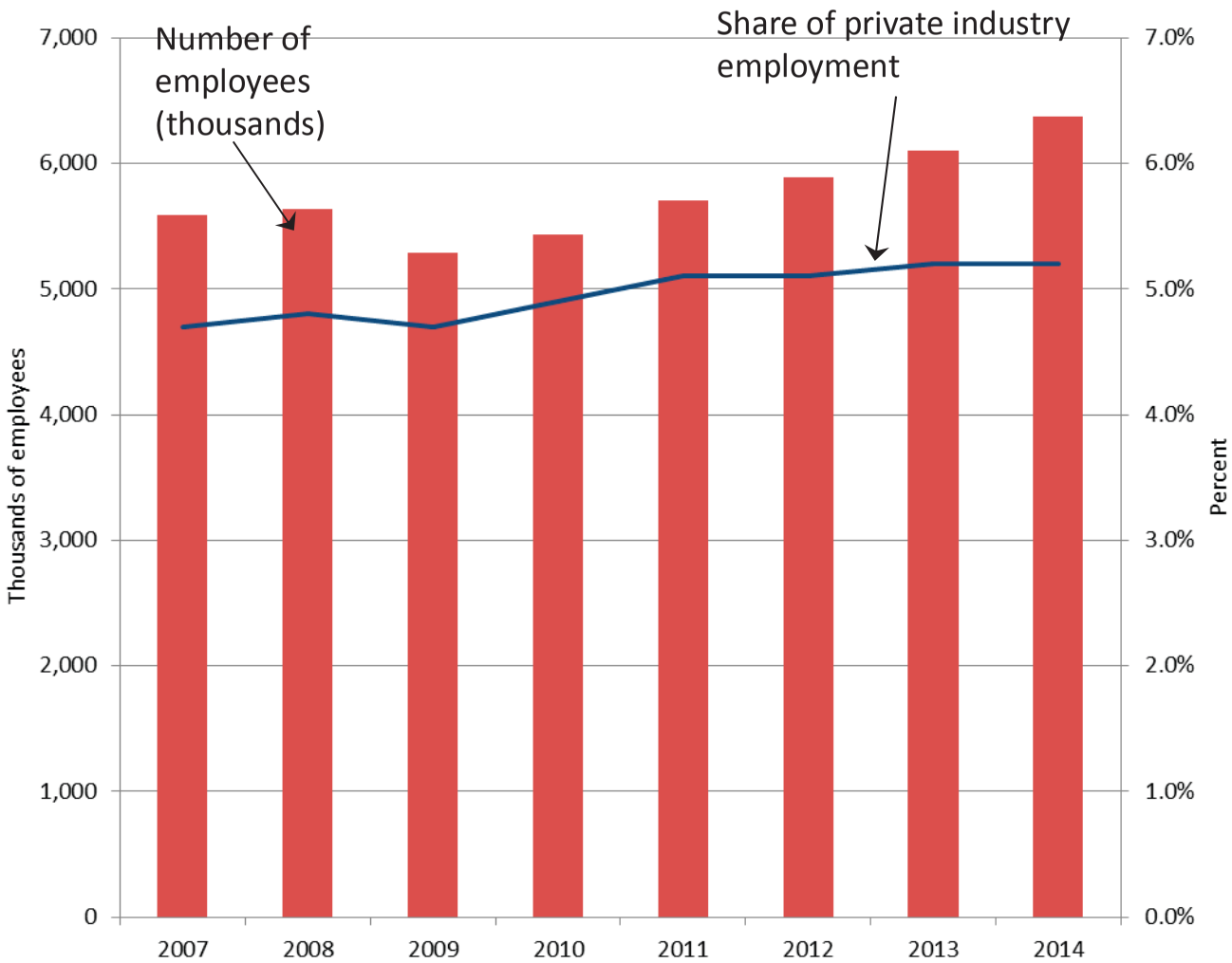
G-20 Workshop on Data Sharing

February 1, 2017

Agenda

- Project linking data on foreign-owned U.S. companies to domestic employment data
  - Data
  - Process
- Benefits of linking
- Challenges
- Future steps

Employment of foreign-owned U.S. companies



Source: BEA

Data on foreign-owned U.S. companies

- Bureau of Economic Analysis (BEA) statistics on Activities of Multinational Enterprises (AMNE)
- Collected by BEA on mandatory benchmark and annual surveys
- Variables collected include:
  - Sales
  - Employment
  - Capital expenditures
  - Financial statements
  - Value added components
  - Exports/imports
  - Research & development
  - Taxes
- Enterprise level

Data on U.S. employment

- Administrative data
- Bureau of Labor Statistics (BLS)
- Quarterly Census of Employment and Wages
- U.S. establishments covered in the unemployment insurance program
- Data collected by states and compiled by BLS

Process of linking data

- Employer Identification Number (EIN)
  - Tax identification number for businesses
  - One company can have many EINs
- Computer match of EINs
- Manual work to link additional establishments to the enterprises
- Use outside sources to identify additional establishments
- Validating the link
  - Over-matched (BLS > BEA)
  - Under-matched (BLS < BEA)
  - Bad matches

Match quality so far

- Work still in progress
- Within 20 percent – “close enough” match

Step in match process	"Close enough" matches	
	Affiliates	Employment
After computerized match of EINs	44.4%	58.7%
After BLS analyst review	52.7%	84.5%
After BEA analyst review	in progress	in progress

Source: BLS

Benefits of linking

- Expand data available for studying effects of direct investment on the U.S. economy
  - More granular detail on FDI employment and wages – industry, geography, occupation
  - Information relating enterprises to establishments – a byproduct of the link – is useful for other linking projects
- Improvement of survey data
  - Linking microdata helps identify errors in survey reporting

Benefits of linking

- Greater frequency of data
  - BEA data are annual; BLS data monthly and quarterly
  - Once initial link is completed, subsequent links may be less labor intensive and could be available with greater frequency
- Potential to reduce respondent burden
  - May be able to reduce data collected on survey if link produces information that meets standards of quality and timeliness

Challenges

- Very labor intensive
  - Substantial investment of time and resources to link data initially
  - Subsequent years may be less work
- Not timely
  - Because of manual effort, substantial delay in producing results
  - Hope that with subsequent matches, data could be more timely
- Legal requirements and limitations
  - Interagency agreement for data sharing
  - Not all states allow BEA to view the data

Future work

- U.S. Government effort to support and expand data linking
  - Employer Data Matching Workgroup
  - 14 Federal agencies represented
  - Report due to be released soon
- Expand collection of identifiers
  - Legal Entity Identifier
  - Over 400,000 entities globally; over 100,000 in the United States
  - BEA plans to collect LEI on upcoming survey
- Pursuing new legislation to expand interagency access to micro-data for statistical purposes

INEXDA

The Granular Data Network

International Network for Exchanging Experience on Statistical Handling of Granular Data (INEXDA)

Stefan Bender (Chair of INEXDA)

INEXDA: The Granular Data Network

On 6th January 2017,

- the Banca d'Italia,
- the Banco de Portugal,
- the Bank of England,
- the Banque de France
- and the Deutsche Bundesbank

have launched INEXDA, an international cooperative project exchanging experiences to declare their willingness to further strengthen their cooperation.

INEXDA
The Granular Data Network

General Mission

- Acknowledging the relevance of micro data in the area of independent scientific research, policy advices (for example monetary policy or financial stability) and statistics and their importance for international comparisons
- Promoting the G20 Data Gaps Initiative II, in particular recommendation 20, addressing the accessibility of granular data
- Acknowledging and supporting the work on data sharing of the Irving Fisher Committee on Central Bank Statistics

INEXDA
The Granular Data Network

Subject and Scope I: Statistical Handling

INEXDA provides a basis for exchanging experiences on the statistical handling of granular data, such as

- the accessibility of data and metadata,
- techniques for statistical analysis of granular data,
- procedures for confidentiality and security of data,
- and methods of output control.

INEXDA
The Granular Data Network

Subject and Scope II: Framework and Aim

INEXDA aims at:

- investigating possibilities to harmonise access procedures and metadata structures,
- developing comparable structures for existing data and
- further fostering efficiency of statistical work with granular data.

The ultimate aim of INEXDA is to facilitate the use of granular data for analytical, research and comparative purposes by users outside the participating institutions, within the limits set by the applicable confidentiality regimes.

INEXDA
The Granular Data Network

First 2 Years: a Pilot Exercise

The five signing central banks have agreed to engage in a pilot exercise which envisages,

1. an extensive stock-taking of available datasets and existing procedures and,
2. the investigation of harmonisation possibilities at different levels.

They will present their results to other interested central banks, national statistical institutes and international institutions by the end of 2018. At the same time, a web page will be launched.

The INEXDA secretariat is provided by the Deutsche Bundesbank for the next two years and can be contacted at INEXDA.secretary@bundesbank.de.

Participation into INEXDA is open to other central banks, national statistical institutes and international organisations.

INEXDA
The Granular Data Network

Thank you for your attention!

- **Contact**: INEXDA.secretary@bundesbank.de

INEXDA
The Granular Data Network

EUROPEAN CENTRAL BANK
EUROSYSTEM

Asier Cornejo Pérez
External Statistics Division
Directorate General Statistics

Data Integration, Linking and Sharing Relevance of International ISO Standards

G-20 Workshop on Data Sharing
Frankfurt am Main, 16 February 2017

Linking datasets

1 Background

2 ECB Experience and initiatives on data integration

3 Cross country data sharing

Increasing data integration needs

- The financial crisis has created a **growing interest in consistent, sound and timely statistics** which implies **larger data volumes** at higher frequency and level of granularity
- In order to **monitor faster economic developments** → **Data integration** from different sources as well as **linking of datasets** is needed together with the **possibility to share information** between institutions and countries
- **Main challenges**:
  - **Data standarisation across the different sources** → Using common agreed International ISO Standards like ISIN and LEI
  - **Confidentiality restrictions for data sharing among institutions and countries** → Using common International ISO Standards and enhancing publicly available information

Micro and granular data systems

- **ECB**, jointly with the European System of Central Banks (**ESCB**), **has promoted the creation of micro and granular data systems** over the last years
- As an example, information on securities and their issuers is compiled in the **Centralised Securities Database (CSDB)** → single information technology infrastructure operated jointly by the ESCB members
  - Data are collected from **various sources**: national central banks, commercial data providers, the public and administrative sources
  - Large data volumes are **automatically compounded** reconciling inconsistent information and detecting incomplete or missing data
    - Only possible due to the use of **common standards** between the sources → In particular, International ISO Standards like the **ISIN, LEI or SNA 2008**
    - Together with **complex algorithms** to overcome the lack of perfect data
  - **Output** → **security-by-security information** with complete and high quality data to the extent possible **shared without restrictions within the ESCB**

Other initiatives and challenges

- In addition to the CSDB, there are other **systems with micro and granular data**
  - Securities Holdings Statistics Database (SHSDB),
  - Money Market Statistical Reporting (MMSR)
  - Register of Institutions and Affiliates Database (RIAD) or
  - In the future, Analytical Credit Dataset (AnaCredit)
- **One key and crucial element** → Use of commonly **agreed International Standards (ISO)** and **publicly available identifiers** like ISIN or LEI
- **Challenges** → Coverage, general application and public availability of identifiers
- **Initiatives** → **ECB Opinions** encouraging the mandatory use of internationally agreed standards as well as the public machine-readable availability of information
  - **Example**: final text of the updated **EU Regulation on the prospectus** to be published when securities are offered to the public or admitted to trading

Use of International ISO Standards

- **Data sharing within the ESCB** is possible without major restrictions, subject to agreed procedures → Also to share information **between EU institutions and countries** in particular areas
- However, there are **plenty of obstacles to provide information with other countries**
  - Technical challenges create impediments for the exchange of information between countries
    - Lack of commonly used identifiers, e.g. for securities and issuers
    - Large data volumes or lack of information for some economies
  - Data sharing challenges - confidentiality or commercial data restrictions

- ***Possible solution*** → *Encourage the use of International Standards (ISO) in all countries and identifiers like LEI, ISIN as a public good to allow sharing of information*

Thanks for your attention

Any questions?

The Usefulness of Common Identifiers and Linking Different Data Sets

Frankfurt, February 1st, 2017

G-20 WORKSHOP ON DATA SHARING

The Usefulness of Common Identifier

Work without system VS work with system

Benefits of Having A System

- enable to meet and surpass all data users expectations
- enable to produce the same quality results every time
- improve performance
- reduce costs
- enable to be an organized organization
- enable to solve all problems on a consistent basis
- enable to be "going global"
- enable to be available anywhere and 24/7

A System is An Enabler

Without System	Benefit	With System
Maybe	Enable to meet and surpass all data users expectations	Yes
Could be	Enable to produce the same quality results every time	Yes
Possibly	Enable to improve performance	Yes
Hopefully	Enable to reduce costs	Yes
Perhaps	Enable to be an organized organization	Yes
Probably	Enable to solve all problems on a consistent basis	Yes
If lucky	Enable to be “going global”	Yes
Should be	Enable to be available anywhere and 24/7	Yes

The Usefulness of Common Identifier

A well managed system requires as follows:

Hardware And Software

Telecommunications

Databases And Data Warehouses

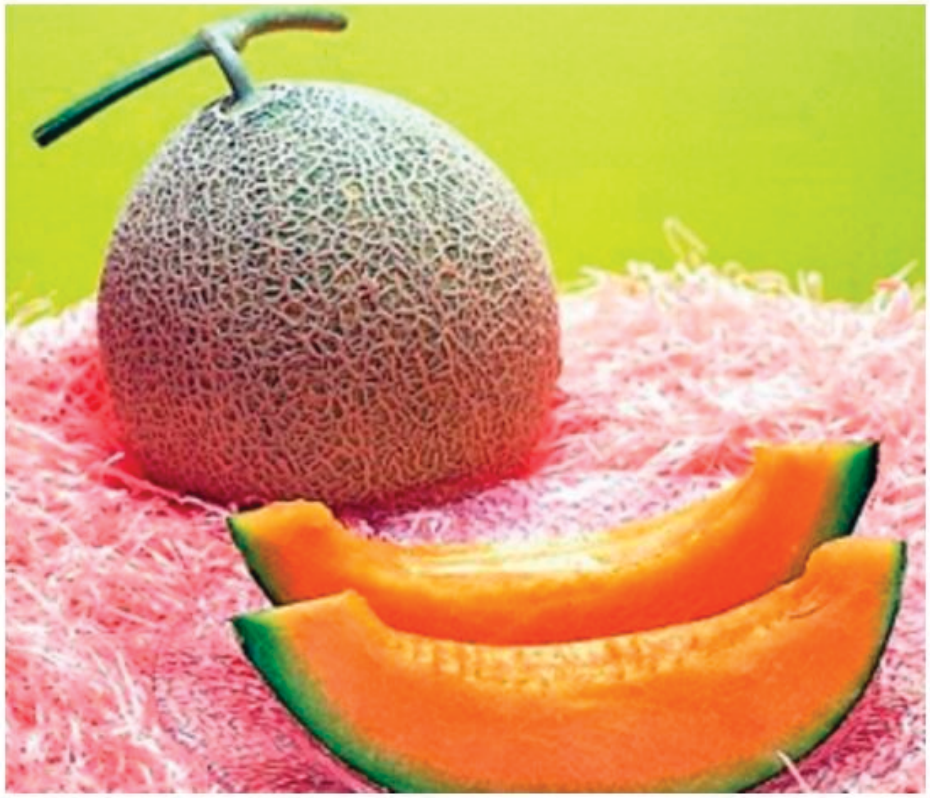
Human Resources

Procedures

Item Identifier

Identifier: Is it relevant here?

#1 – Yubari Melon ($23,000/Per Pair)



Implementation of Identifier depends on the value of the item:

Melons → No

Handphones → Yes

Cars → Yes

Very special melons that cost 11.500 USD each → Yes

so, depends on the **value** of the item

The Usefulness of Common Identifier

As a Valuable Item in the system, micro data package has to be identified with an identifier

Reasons that make micro data valuable:

- Maximum granularity
- Involving public convenience and security

The Usefulness of Common Identifier

A unique identifier will be assigned on every micro data package

Q: Possible?

A: With current state of IT → Very Possible

Example of massive handling by current IT system: transaction code, booking code, etc.

Key Principles Underlie the Identifier:

1. It is a global standard.
2. A single, unique identifier is assigned to each data package. Requires concept maturity and coordination among country coordinators (local operating units).
3. It is supported by high data quality.
4. Coherent with other recommended identifier (e.g. ISO 17442 LEI, learning from UPI, UTI, USI). Coherent does not mean similar. But adopt the equal standard.
5. Supported by good IT system that can cope with high speed transaction and database that can accommodate massive and highly dynamic content.
6. All involved elements are bound in a system. Each element has liabilities and benefits (non financial and financial).

Learning from LEI

"The founding principles of the Global LEI System were developed through extensive public and private sector collaboration and will continue to evolve in this spirit. At their Cannes Summit in November 2011, the G-20 leaders supported "the creation of a global legal entity identifier (LEI) which **uniquely identifies parties to financial transactions**." The leaders also called on the Financial Stability Board (FSB) to take the lead in helping coordinate work among the regulatory community on the governance framework of the Global LEI System, complementing efforts by the private sector to develop a technical solution, including through the International Organisation for Standardisation."

Learning From Other Identifiers

UPI (Unique Product Identifier), A unique code to describe a financial product for the purpose of regulatory reporting.

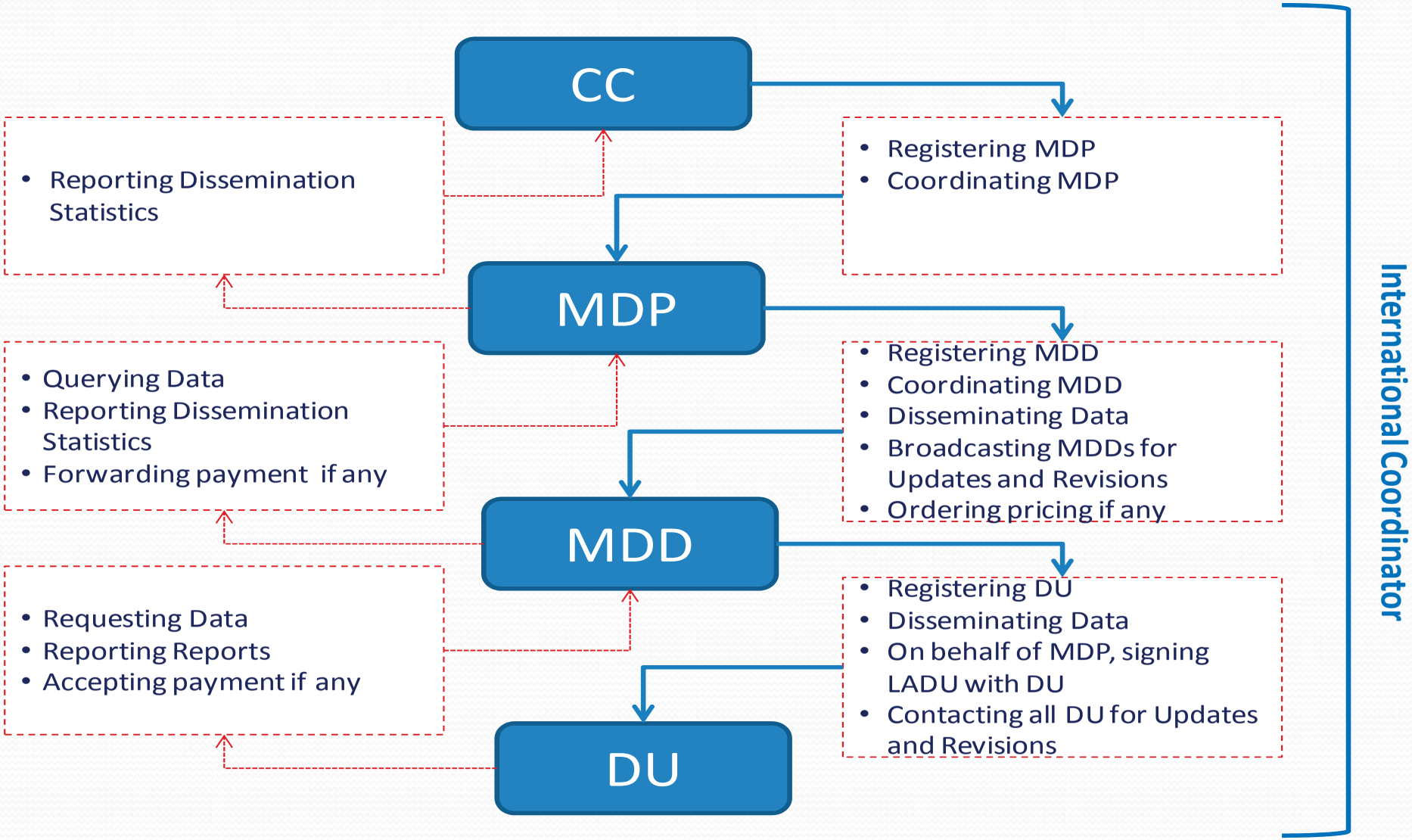
UTI (Unique Transaction Identifier), With a LEI and a UPI embedded in a financial transaction, the counterparty and the product traded can be known. Now the financial transaction itself needs to be identified, as there can be many of the same transactions conducted by the same parties in the same product.

FEI (Financial Event Identifier), identifying a financial event such as a merger, acquisition, bankruptcy, spin-off, etc

The integrated global identification system, Promotes U3 (Unique, Unambiguous, and Universal) Global Identification coding scheme. It conforms to all known standards to date (ISO LEI 17442:2012) and to further expected UTI and UPI requirements.

The Usefulness of Common Identifier

Proposed Structure As The Environment of Micro Data Package Identifier



Proposed Format of Micro Data Package Identifier

As a draft idea, Micro Data Package Identifier at least consists of 22 digit of codes as follows; Country (or International Organization) code (e.g. ISO 3166-2 alpha-3) [Digit 1-4], Micro Data Provider (MDP) code [Digit 5-7], Micro Data Distributor (MDD) code [Digit 8-10], Dataset code [Digit 11-13], Unique incremental number for each MDD[Digit 14-16], Data User code [Digit 17-19], Unique incremental number for each DU request [Digit 20-22].

Linking Different Data Sets

Linking Different Data Sets would significantly enrich the information provided by data sets

It Consists of:

- Variable/Field Linking
- Coverage Linking
- Time Series Linking

Benefits of the Ability of Data Linking

On Data: Improves coherency and consistency.

On Data User: Possibility to have more comprehensive data.

On Data Provider: Efficiency in data management and processing, Improves data quality.

Requirements for Linking

* **Availability of Metadata**
  → Similar perception on the data
  →Avoiding error of acceptance or rejection

* **Eligibility Analysis of Data Linking** prior to linking process

Eligibility Analysis of Data Linking

Eligibility Analysis is performed **in the background**. It starts to run as soon as the data package is available for a distribution as well as metadata.

To enable the widest possible of data linking, the eligibility analysis is done by **Micro Data International Coordinator**.

Challenges in Data Linking

Variable/Field Linking: requires precision on record profiling to avoid mismatch. Beware! The more linked, the more complete information on a row. It means more possibility of identity reveal.

Coverage Linking: duplication and under coverage are the common mistakes. Requires also precision on record profiling and enough underlying information on the datasets involved.

Time Series Linking: Has to be aware of changes in data structure along the period.

The Role of Identifier in Linking Data Sets

Significantly speeding up the process of addressing data sets to be linked

- Locating the data
- Determining which data parts
- Determining the time reference of data sets

Improving accuracy of data processing by ensuring row identity

Coupled with good metadata, will enable the creation of global interlinked data

Improving process standardization that will reduce the possibility of leaks on individual data

Thank You!

BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

DATA SHARING: EXPERIENCE & CHALLENGES OF INDONESIA'S STATISTIC

Gantiah Wuryandani, Etika Rosanti

G-20 Workshop on Data Sharing
Frankfurt, January 31- February 1, 2017

BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

OUTLINE

Introduction

Concept, purpose and benefit of data sharing

Current practice in Indonesia

Obstacles in data sharing

Major challenge/issues

BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

INTRODUCTION

- The need of data sharing became more crucial for Indonesia, particularly since economic crisis in 1997/98.
- Policymakers must have comprehensive information and clear picture of issues.
- Holistic analysis support pinpoint the right problem and its solution.
- The need of more granular and transparent data to support holistic analysis.

BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

CONCEPT, PURPOSE AND BENEFIT OF DATA SHARING

Concept

- Data sharing is the practice of data exchange between various organizations, people and technologies.
- Data sharing is lawful, and confidentiality should be maintained.
- Data sharing is usually followed by risk, either perceived or actual. Therefore, it needs risk management to avoid the potential risk in misused data.

BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

CONCEPT, PURPOSE AND BENEFIT OF DATA SHARING

Purpose

- To have synergy with other institution
- Improve efficiency and transparency
- To have better policy formulation and its effectiveness.
- More granular and transparent data is aimed to educate public and smooth over the establishment of specific policy.

CONCEPT, PURPOSE AND BENEFIT OF DATA SHARING

Benefit

- Wide ranging information for all policymakers
- Clearer and better quality of data
- Minimized burden of redundant data collection effort, particularly reporting from the same respondent entities by different authorities
- Efficient in data collection

CURRENT PRACTICE

Legal Basis

- **Act No. 23 of 1999 regarding Central Bank**, article 14 paragraph (1) states that central bank may conduct a survey regularly either at macro or micro level to support the functions in monetary, payment system, and financial stability policies
- **Act No. 24 of 1999 regarding the Foreign Exchange Flows and Exchange Rate System**, article 3, conveys to BI the authority to request information and data of foreign exchange transactions, assets and liabilities of any institution
- **Central Bank Regulation on Reporting** particularly for banks and non financial institution

7

CURRENT DATA SHARING

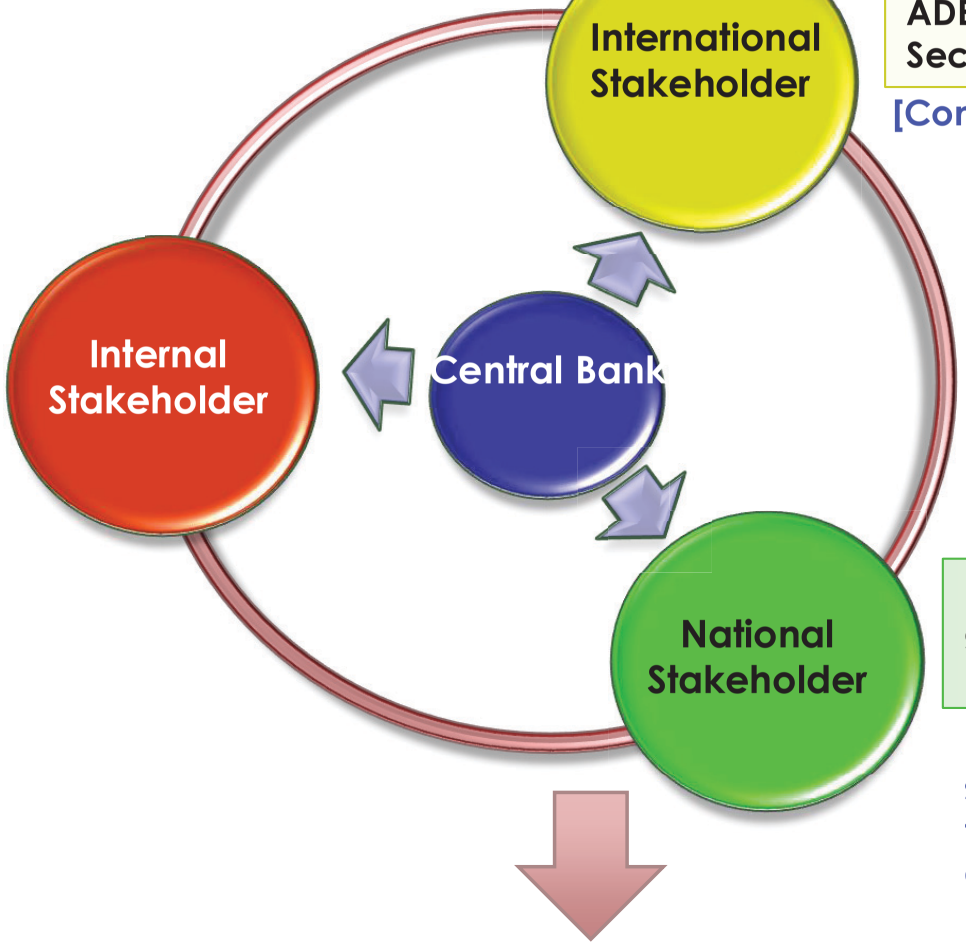
International

- Submission to IMF (SDDS, IFS)
- Commitment realization in DGI G-20
- Submission to BIS, ADB, OECD, World Bank, ASEAN Secretariat

National

- Integrated system for data sharing with FSA
- Joint compilation with NSO: as contributor in GDP, Flow of Fund, business survey, research in WPI
- Join banking reporting maintenance and utilization with FSA
- Join compilation with MoF in debt securities and foreign loan
- As the government bond custodian and settlement
- Join export - import data sharing system with NSO, Custom, and Tax Office

DATA SHARING FRAMEWORK



BIS, IMF, World bank, ADB, DGI, ASEAN Secretary, OECD, etc.

[Commitment base]

Board of Governor, other Dept. & regional office

[Legal base: internal regulation, special access]

NSO, MoF, FSA, local government, other ministries/ institutions

[Legal base: MoU, special access for FSA, integrated system of data sharing]

Way Forward : "ONE DATA FOR NATIONAL"

EFFORT TO STRENGTHEN DATA SHARING

- Promote MOU with other institutions as data sources
- Develop integrated exchange system with Financial Supervisory Authority
- Improve granular and transparency of statistical publication in order to escalate institutional sectors alertness of data as the early warning condition and to prevent instability by establishing preemptive policy
- Regular exchange in information through focus group discussion and data exchanges with other institutions
- Maintain confidentiality by regular review and assessment on access provided
- Adopt international standard best practice methodology in statistic compilation
- Participate in International statistics commitment such as IMF (IFS, SDDS), DGI-G20, BIS, OECD, ADB, World Bank, ASEAN Secretariat
- Join compiling in some indicators such as Sectoral Accounts, flow of fund, GDP, debt securities, external loan with other institutions (NSO, MoF, FSA and local government).
- Way forward, establish an integrated reporting system in banking industry (collaboration of central bank and FSA)

FEATURES OF MOU

Purpose

Increasing the commitment, cooperation, and synergy in order to exchange data/statistic information related to monetary policy, payment system, financial system stability, real sector and other particular coverage, and also developing the human resources capabilities and research.

Target data to be shared

Data related to specific needs between central bank and other institutions bilaterally or multilaterally agreed on. Exclude individual data as this is prohibited by law for banking data, restricted and confidentiality data, unless stated to be shared.

Procedure for data sharing

- The format, mechanism, and process of data sharing with other institution should be established.
- Data exchange beyond agreed, should be requested by special letter
- Inclusion of discussion forum for building capabilities and research in the MOU
- Confidentiality of limited data sharing

FEATURES OF MOU

Limitation on use

The information provided should not be used for other purposes than the institution

Transferring data

Information is exchanged via media agreed by both institutions.

Protect confidentiality

stipulate ways to protect the confidentiality of information :

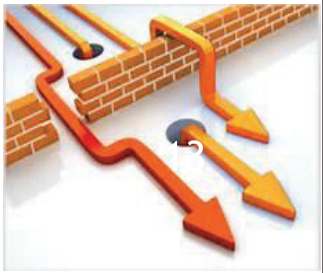
- Approval of data-sharing access be governed by reviewing the candidate's position and task
- Appropriate actions to ensure non-disclosure of confidential information, including sanction
- Regulations and code of conduct for persons who have access to confidential information from misused

Mediation of disagreement:

All disputes/disagreement among the institutions will be resolved by consensus based on regulation.

OBSTACLES IN DATA SHARING

- Legal and confidentiality constraint
- Protection of confidential information in term of legal foundation (data security)
- Standardize methodology and metadata
- IT system and cost
- Accessibility: open or restricted, code of conduct to safeguard the security
- Fear of strategic management hampering
- Difficulties in data collection, particularly in corporation and household sectors
- Institutions reluctance to open access for others



MAJOR CHALLENGES/ISSUE

Protocol to secure data confidentiality (sensitive/individual/personal information)

Standardized metadata and methodology as to have reliable analysis in countries comparison of cross border transactions

Data clearing and reconciliation among countries cross border statistics, and the need of statistics compilers mailing list/contact.

Integrated system in join compiling statistics

Effectiveness of data sharing utilization to map instability of cross border transaction/position to prevent crisis and safeguard stability by join intervention among countries.

Data reliability in the cross fertilization as the mirroring data for countries with missing data of outward flows

Thank You
Danke

DEUTSCHE BUNDESBANK
EUROSYSTEM

Record Linkage

Definition and German Experience

Stefan Bender (Head of the Research Data and Service Centre), Deutsche Bundesbank

Motivation: The Need for Record Linkage

- **Large amounts of data** are being collected.
- Increase **analytical value** of data
- Improve **data quality**
- Reduce **survey burden** for units (like companies, banks)
- Data are often from **different sources** (need for record linkage).

Definition of Record Linkage

- RL is finding records in different data sets that represent the same entity and link them.

- RL is also known as *data matching, entity resolution, object identification, duplicate detection, identity uncertainty, merge-purge.*

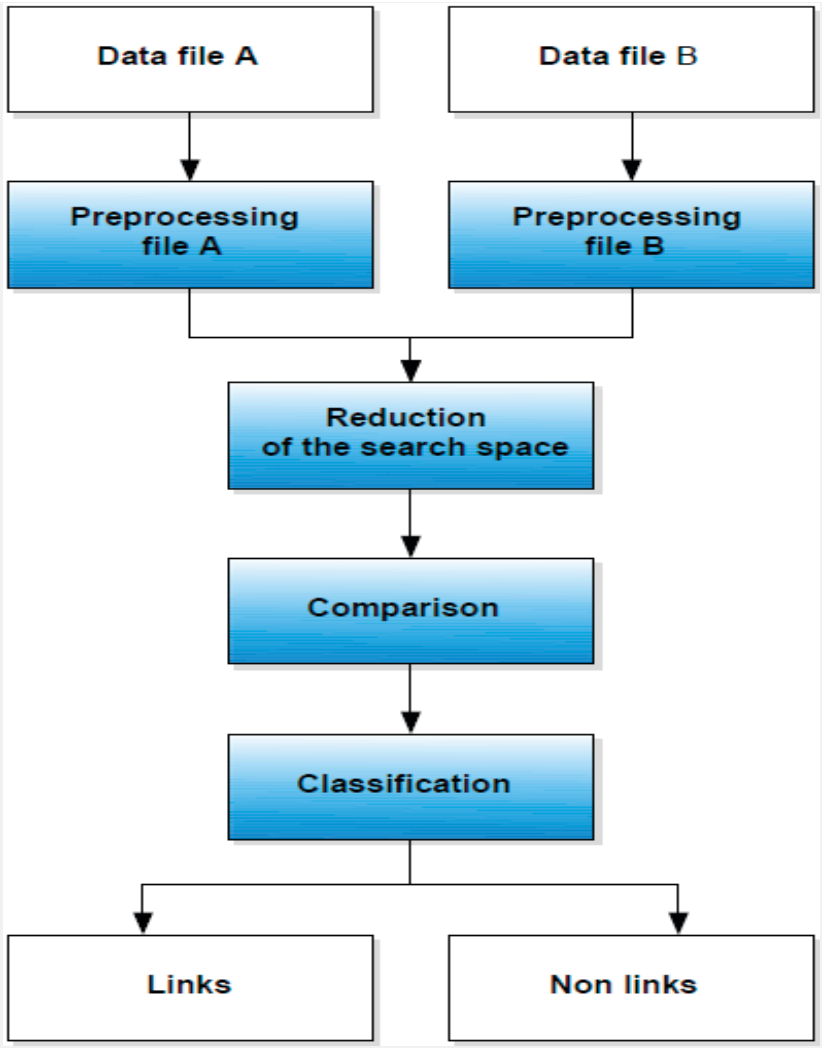
Main Applications of Record Linkage

1. Merging of two or more data files
2. Identifying the intersection of the two data sets
3. Updating of data files (with the data row of the other data files)
4. Impute missing data
5. Deduplicate a file (remove duplicates in one file)

Record Linkage Challenges (Christen 2012)

- No unique (clean) entity identifiers available
- Real world data are dirty (typographical errors and variations, missing and out-of-date values, different coding schemes, etc.)
- Scalability, data base size
  - Naïve comparison of all record pairs is quadratic
  - Remove likely no-matches as efficiently as possible
- No training data in many linkage applications
  - No record pairs with known true match status
- Privacy and confidentiality
  - Personal information, like names and addresses, are commonly required for linking

The extended record linkage process



There is no perfect world

- In a perfect world

		Predicted 1	Predicted 0
True	1	True Positive (TP)	
	0		True Negative (TN)

- But we do not live in a perfect world

		Predicted 1	Predicted 0
True	1	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

Record Linkage at Bundesbank (in the RDSC)

Background

- **No common register** in the Bundesbank
- Strong need to have an integrated company register
- **Use of 7 different data sets** (internal and external) to construct some kind of an integrated company register (for one year):
  - companies from foreign direct investments (MiDi),
  - balance of payment statistics (SITS),
  - banking supervision data on borrowers (BAKIS / MiMiK),
  - balance sheet data (USTAN) and
  - external balance sheet data

Challenge

- **No common unique firm identifier** in Germany
  (Company business register-ID **not stable**)

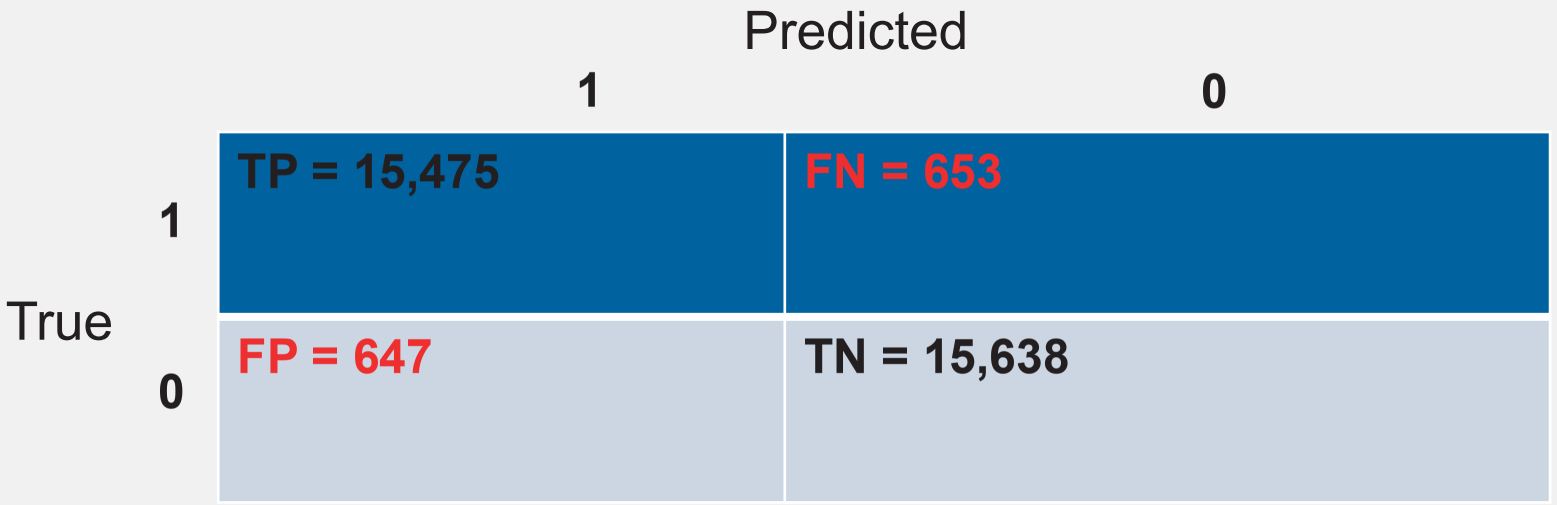
Record Linkage at Bundesbank (in RDSC)

- Match firm data…
  - … that do not have a common unique identifier / key
  - … by using alternative identifiers (such as names)

- It is possible to construct a „ground-truth“-sample of match candidate pairs to train a RL model:
  - Common external Ids
  - Quasi identical balance sheets of firms

- Machine Learning Algorithm, which can be used for other years or company data sets.

Result of two Company Data Sets



- Results of RL are very good and will be used for other linkage.

Record Linkage across Institutions: KombiFiD

In the project Combined firm data for Germany (KombiFiD) data collected by

- German Statistical Offices,
- Federal Employment Agency, and
- Deutsche Bundesbank

were for the first time linked.

With a huge effort - after 5 years (2007-2011) and a lot of persons involved - the data were linked and made available to the research community. Most efforts in the following tasks:

- Legal issues
- Cleaning
- Record Linkage
- Consistency checks

Costs

- Linking these data are costly
  - High effort to clean the data (most time is spend in cleaning)
  - Uncertainty of having all true matches
  - Legal „uncertainties”

- Strong need for better data quality.
- Stronger need for unique identifiers like LEI

Thank you for your attention!

- **Website**: www.bundesbank.de\fdsz
- **Contact**: fdsz@bundesbank.de