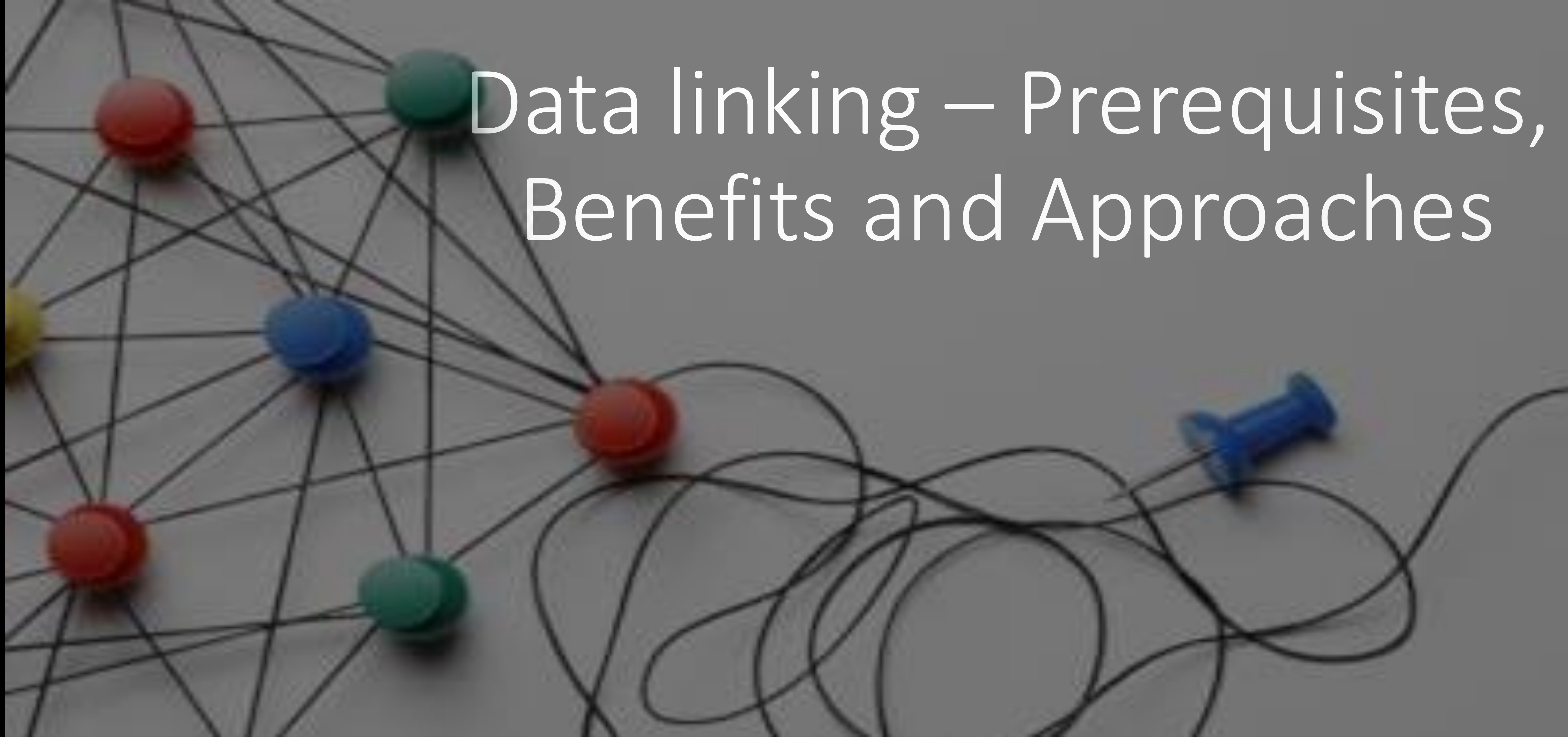
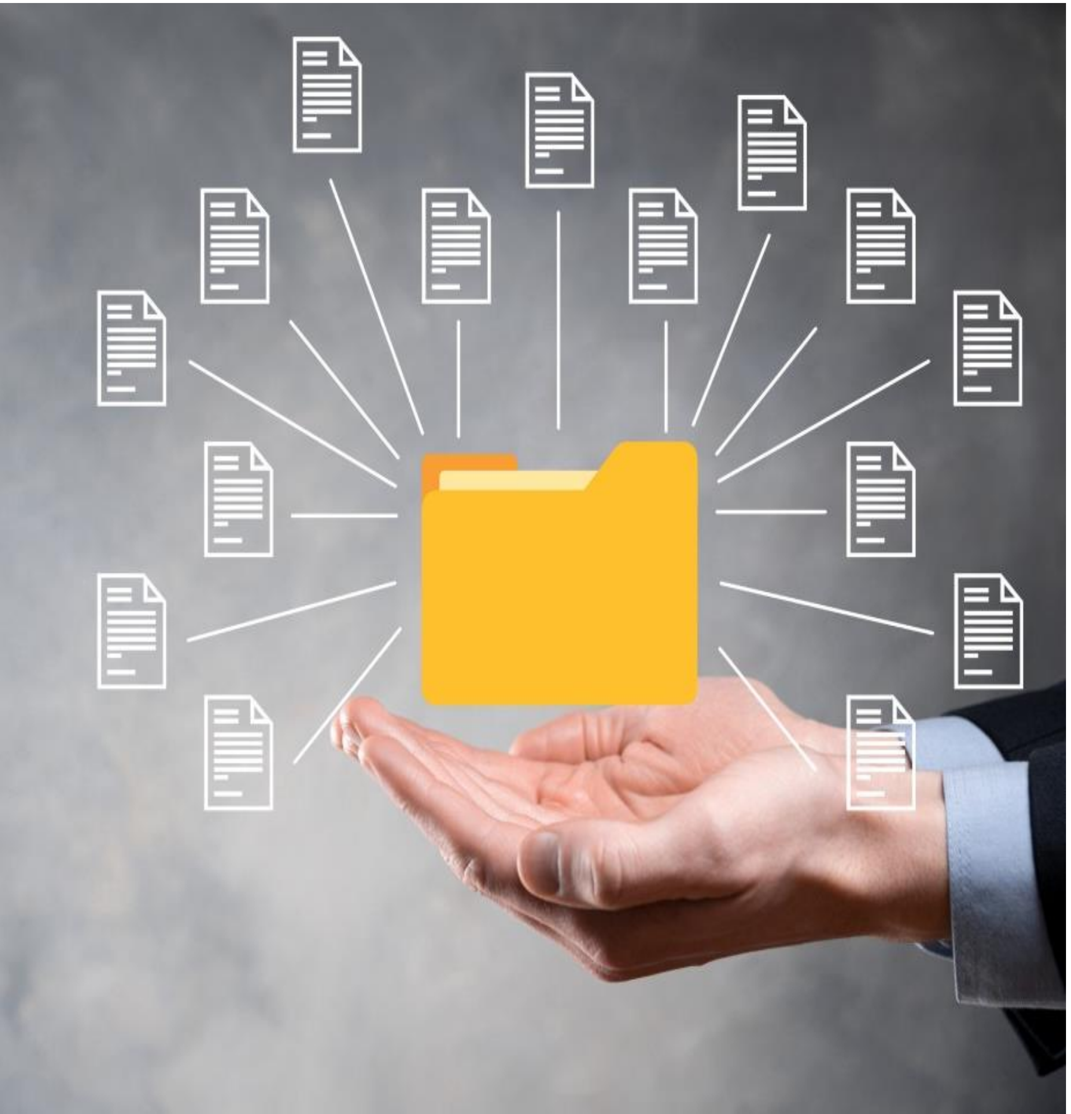


Data linking – Prerequisites, Benefits and Approaches

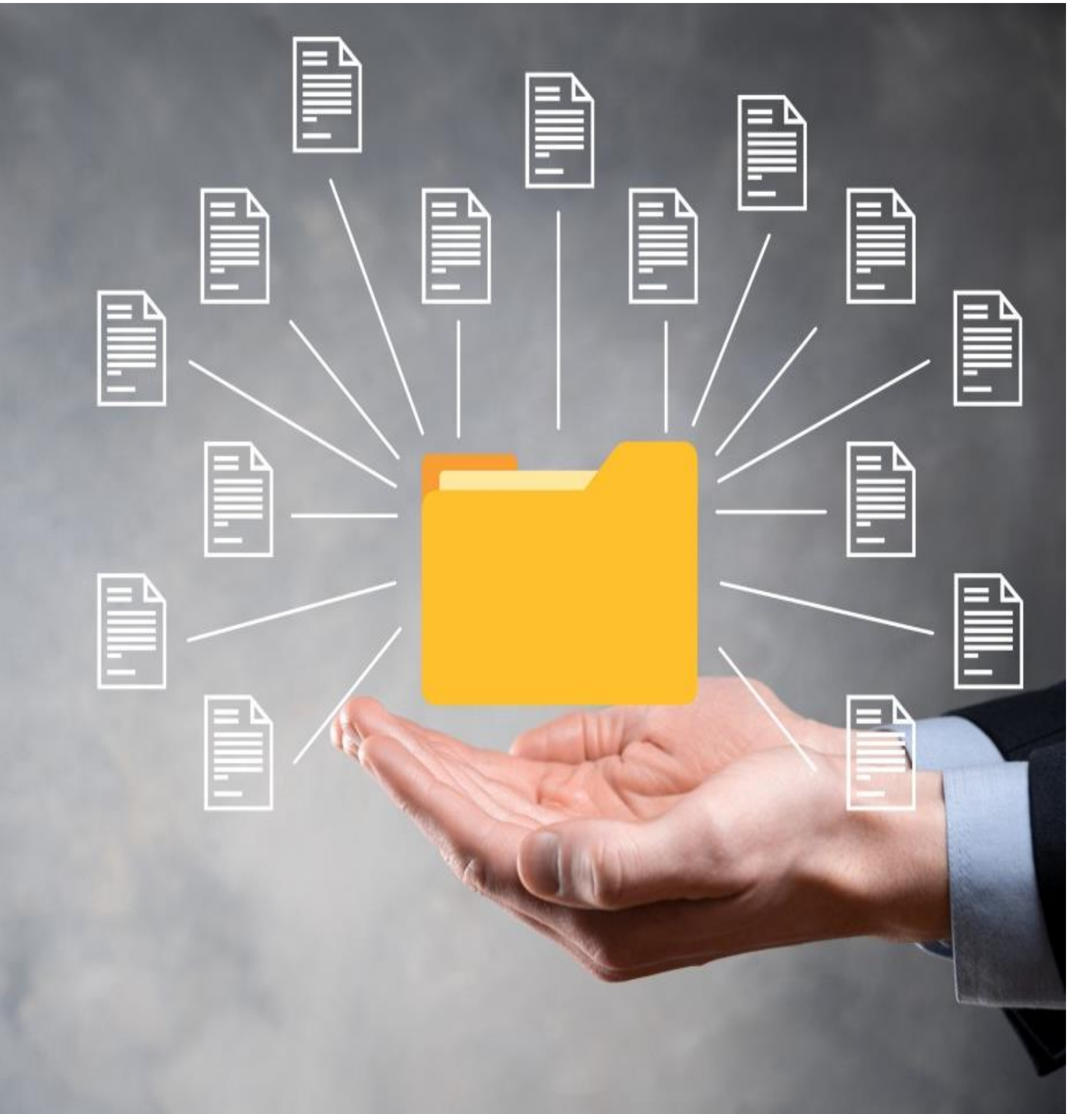
United Nations | DESA Statistics Division

What is data linking?



Data linking is the **process of identifying, matching and merging records** that correspond to the same entity (person, firm, other) from several datasets or even within one dataset.

How do we link?



There is need for a “key” to link or merge records referring to the same unit (ex: same person, household, firm, etc.)

Datasets are linked using identifiers

- Unique Identifiers; often a number
- Other identifiers that are not necessarily unique; name, address, gender, date of birth, a specific ID number, or a combination of identifiers, etc.

Why do we want to link data?

We are in a context of increasing availability of data from different sources with high dispersion in terms of their quality and in general their characteristics.

- Why do you think data linking is important?
- What are key aspects to think of when linking data?

Why do we want to link data?

- Sometimes, different datasets provide for opposing conclusion regarding a socioeconomical phenomenon, thus creating tensions at policy maker levels and mistrust from civil society.
- In other cases, different data sources may be incomplete, or provide only parts of the information. Data integration also in these cases is important
- To improve evidence-based decision making, there is a need to integrate those data sources, to improve their quality and with that our understanding of specific phenomenon and how to address them

The importance of individual records

Databases are not always “ready to be linked”, but instead its linking poses a set of challenges:

- Databases could have records with errors (typo, outliers) or being incomplete (missing values), so a preliminary check of basic quality conditions and some editing must be performed prior to any merge.
- Likewise, two variables from different datasets could have the same name but different definitions, so some prior standardization must be ensured as well.
- For increasing volume of data, there is also a challenge regarding the computational complexity.

Unique Identifier

- A unique identifier is an identifier that is guaranteed to be unique among all identifiers used for those objects and for a specific purpose.
- Unique identifiers can take different forms:
  - Serial numbers
  - Random numbers, codes or names
  - Number, code or name that contains personal information

Example: Norway

- All persons in Norway have a unique personal ID.

$$\underbrace{120572}_{\text{Birth date}}\ \underbrace{3492}_{\text{Random number}}\underbrace{3}_{\text{Gender}}$$

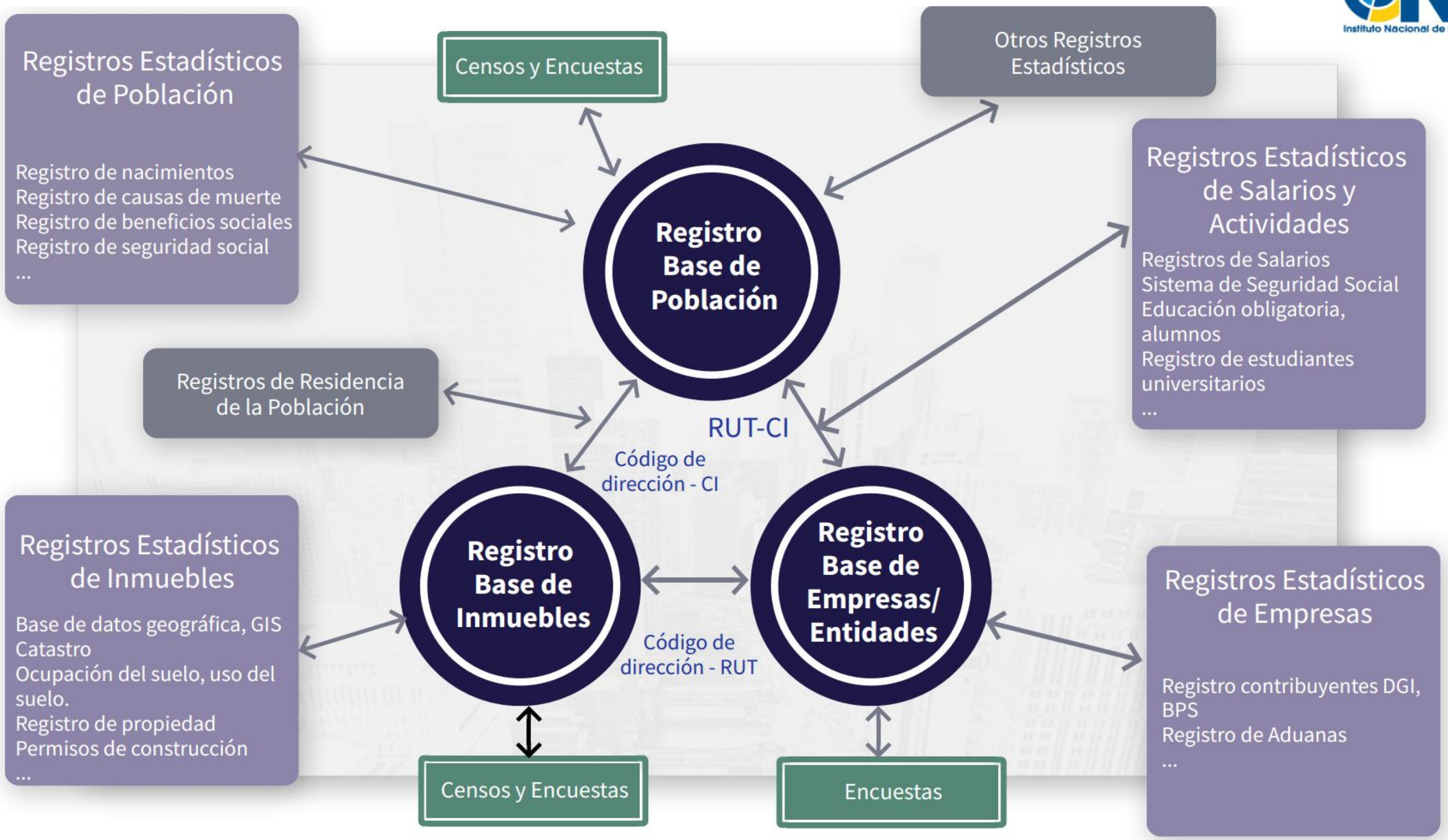
Challenges:

- Running out of numbers for certain dates
- Incorrectly entered birth date
- Change of gender

➡ Norway recommends to use randomly assigned numbers to avoid these challenges

The value of data linking: at economy-wide data level

Integrated System of Statistical Registries and surveys (SIREE – Uruguay)



Registros Estadísticos de Población

Registro de nacimientos
Registro de causas de muerte
Registro de beneficios sociales
Registro de seguridad social
...

Censos y Encuestas

Otros Registros Estadísticos

Registro Base de Población

Registros Estadísticos de Salarios y Actividades

Registros de Salarios
Sistema de Seguridad Social
Educación obligatoria, alumnos
Registro de estudiantes universitarios
...

Registros de Residencia de la Población

RUT-CI

Código de dirección - CI

Registros Estadísticos de Inmuebles

Base de datos geográfica, GIS
Catastro
Ocupación del suelo, uso del suelo.
Registro de propiedad
Permisos de construcción
...

Registro Base de Inmuebles

Registro Base de Empresas/ Entidades

Código de dirección - RUT

Registros Estadísticos de Empresas

Registro contribuyentes DGI, BPS
Registro de Aduanas
...

Censos y Encuestas

Encuestas

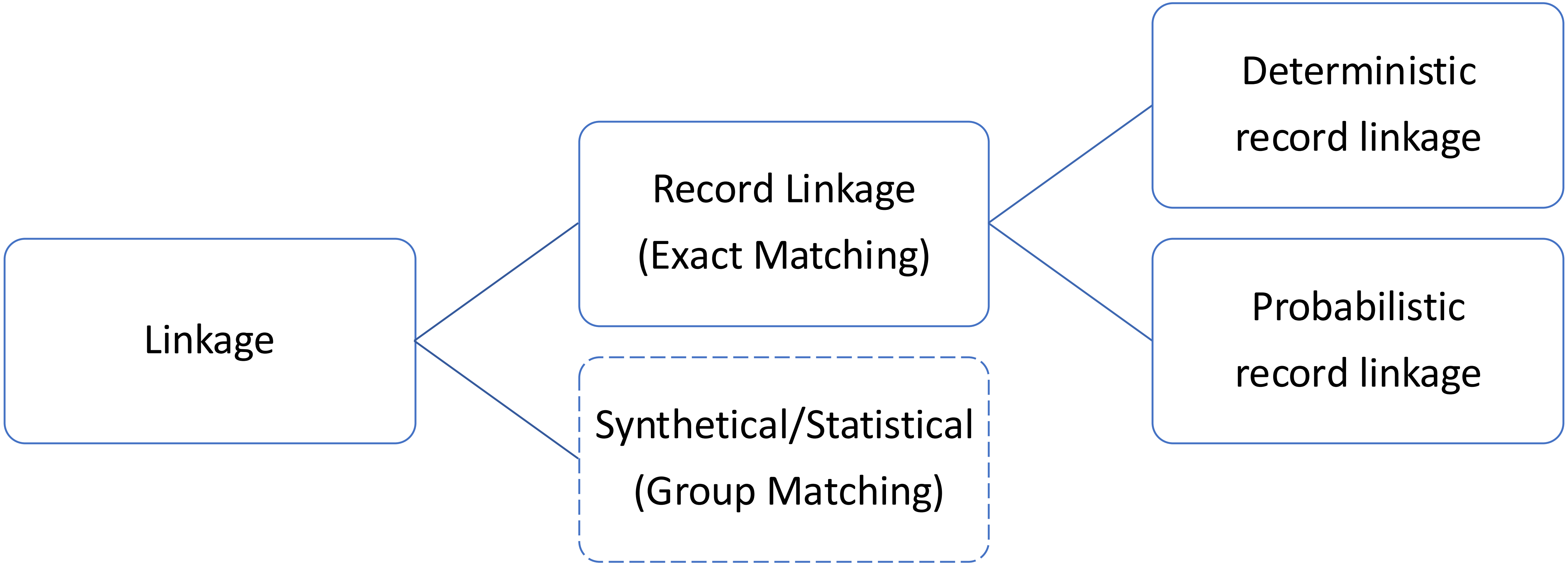
The NSO of Uruguay implemented a system of linked registries (population, dwelling, economic entities) in order to be able to conduct a pilot Census based on admin data by 2023, and a definitive Census in 2030, based on admin data and/or mixed modality.

Data Linkage at technical level

- **Record linkage** refers to the identification and combination of records corresponding to the same entities – persons, enterprises, dwellings, households, etc. – throughout two or more data sources. There are two standard ways this is done. **Deterministic** and **Probabilistic** approach.

- **Statistical matching** refers to combining data that may not necessarily correspond to same entity as person or business but refer to same population group. It is mainly used in market research and **less used** in NSO/NSS context.

Types of data linkage



Linkage

Record Linkage
(Exact Matching)

Synthetical/Statistical
(Group Matching)

Deterministic
record linkage

Probabilistic
record linkage

Deterministic record linkage

- It’s the simple record linkage between datasets based on unique identifiers like social security number, national ID, goecode of address, or similar variables among the available data sources. It could also be a combination of variables to generate identifier to link records.
- A possible difficulty is that unique identifiers could also be affected by errors occurred for instance during either the data collection, or the data capture processes.

Probabilistic record linkage

- When there is no unique identifiers, we can use a probabilistic approach where a score is calculated based on selected variables like name, data of birth, address, etc. for each potential pair linking (mostly based on the Fellegi and Sunter model 1969)
- These are more prone to be affected by data collection or data capture errors, or they are often recorded in different formats making their comparison more complicated.

Typical steps in probabilistic linking

- **Preprocessing** (Standardizing and cleaning the data to ensure consistency)
- **Blocking** (Grouping records that share similar characteristics to reduce the number of record pairs)
- **Comparison** (comparison methods like string matching algorithms or distance metrics)
- **Classification** (A threshold is often set to decide which pairs are considered matches)
- **Evaluation and refinement** (through validation processes and refining the model or parameters

Example

ID	Name
1000323	Paul Verbena
1003723	Elisa Becket
1004943	John Castle
1005396	Jessie Jackson

ID	Name	Birth date
1000323	Paul Vervena	05.02.1973
1002723	Elisa Becket	04.07.1947
1004943	John Castle	07.08.1964
1005396	Jeremy Jackson	12.06.1985

Name	Birth date
Paul Verbena	02.05.1973
Elisa Beckett	04.07.1947
Jon Castle	07.08.1964
Jessie Jackson	12.06.1985

Example: deterministic linkage

ID	Name
1000323	Paul Verbena
1003723	Elisa Becket
1004943	**John Castle**
1005396	Jessie Jackson

ID	Name	Birth date
1000323	Paul Vervena	05.02.1973
1002723	**Elisa Becket**	**04.07.1947**
1004943	**John Castle**	**07.08.1964**
1005396	Jeremy Jackson	12.06.1985

Name	Birth date
Paul Verbena	02.05.1973
Elisa Beckett	**04.07.1947**
Jon Castle	07.08.1964
Jessie Jackson	12.06.1985

Example: probabilistic linkage

ID	Name
1000323	Paul Verbena
1003723	Elisa Becket
1004943	John Castle
1005396	Jessie Jackson

ID	Name	Birth date
1000323	Paul Vervena	05.02.1973
1002723	Elisa Becket	04.07.1947
1004943	John Castle	07.08.1964
1005396	Jeremy Jackson	12.06.1985

Name	Birth date
Paul Verbena	02.05.1973
Elisa Beckett	04.07.1947
Jon Castle	07.08.1964
Jessie Jackson	12.06.1985

Some tools for probabilistic record linkage

- **AutoMatch**, developed at the US Bureau of Census, now under the purview of IBM [Herzog et al. 2007, chap.19].
- **Febrl** - Freely Extensible Biomedical Record Linkage, developed at the Australian National University [FEBRL].
- Generalized Record Linkage System (**GRLS**), developed at Statistics Canada [Herzog et al. 2007, chap.19].
- **LinkageWiz**, commercial software [LINKAGEWIZ].
- **RELAIS**, developed at ISTAT [RELAIS].
- **DataFlux**, commercialized by SAS [DATAFLUX].
- **The Link King**, commercial software [LINKKING].
- **Trillium**, commercial software [TRILLIUM].
- **Link Plus**, developed at the U.S. Centre for Disease Control and Prevention (CDC), Cancer Division [LINKPLUS].

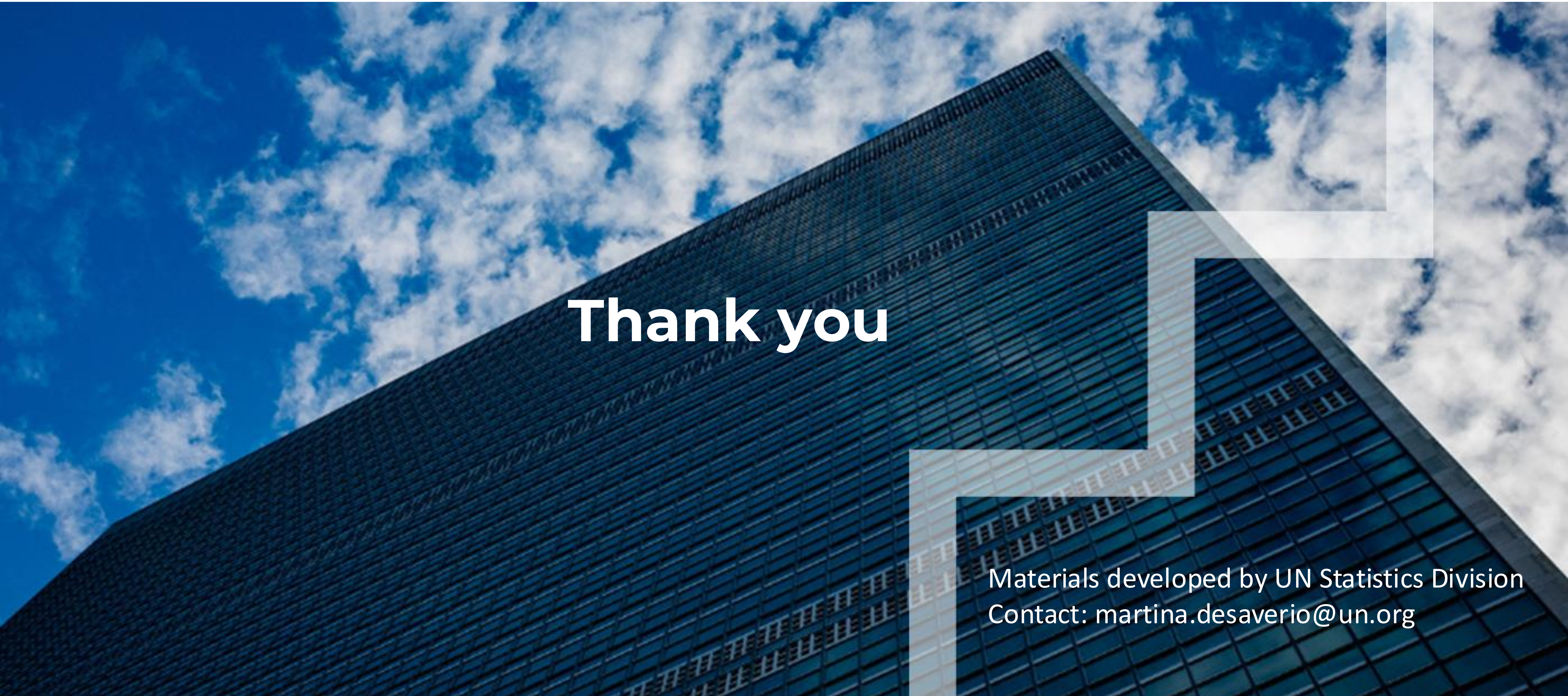
Some libraries for deterministic/probabilistic record linkage

Python	R	Java
Atylmo FEBRL FuzzyMatcher Python record linkage toolkit RLTK Splink Zingg hlink	fastLink RecordLinkage Reclin2	FRIL JedAI

Useful links and references for data linking

- Winkler, William. (2014). Matching And Record Linkage. Wiley Interdisciplinary Reviews: Computational Statistics. 6. 10.1002/wics.1317.
- UNECE - A Guide to Data Integration for Official Statistics
- Report of WP3. Software tools for integration methodologies
- Stats Canada – Power from data 2021
- Open source and/or freely available data matching software
- https://unstats.un.org/UNSDWebsite/capacity-development/admin-data/detailedResourceView/resource_clinic_probabilistic_record_linkage

United Nations | DESA Statistics Division



Thank you

Materials developed by UN Statistics Division
Contact: martina.desaverio@un.org