Combining microdata: the art of matching and joining

SSD

SSD System of Social Statistical Datasets

Egon Gerards

The challenges of combining data

Matching; assigning linkage keys

Joining: the SSD

SSD

SSD System of
Social Statistical Datasets

Matching: personal identifiers

Datasets contain personal identifiers

1. Direct unique personal identifiers:

- &Ł–Łû˘ı#6˘ßóŁ″˘#ł °æ ¸˘ß
- ’ ˚–˘#ł–#¸Łß–ª
- 1 ˚æ ˘
- $ ˛ ˛ß˘¹¹
- 6˘ö
- 3ł¹–˚°#″ł ˛˘
- (–″

2. Indirect/other personal identifiers

Matching: personal identifiers

1. Direct unique personal identifiers:

--> replaced by SN linkage keys or removed/adjusted

- PIN: person identification number
- AIN: address identification number

2. Indirect/other personal identifiers:

- ß˘æ łó˘ ˛#
- ˘ı ˝ßÊø–˘ ˛

Matching: population register

The Dutch population register is the key factor in matching!

- Cumulative data from 1995 onwards
- Every month updated

Personal identifiers	Unknown (1995 - 2015)
Citizen Service Number (CS) – BSN	<0,0025 %
Date of birth	PR contains no unknown, but 1st january and 1st july are often used when exact date is unknown.
Sex	0%
Postal code	<0,09%
Number of the house	<0,09%

Matching: strategy

Matching strategy

1. Optimizing number of matches
2. Minimizing number of mismatches and missed matches

Matching keys

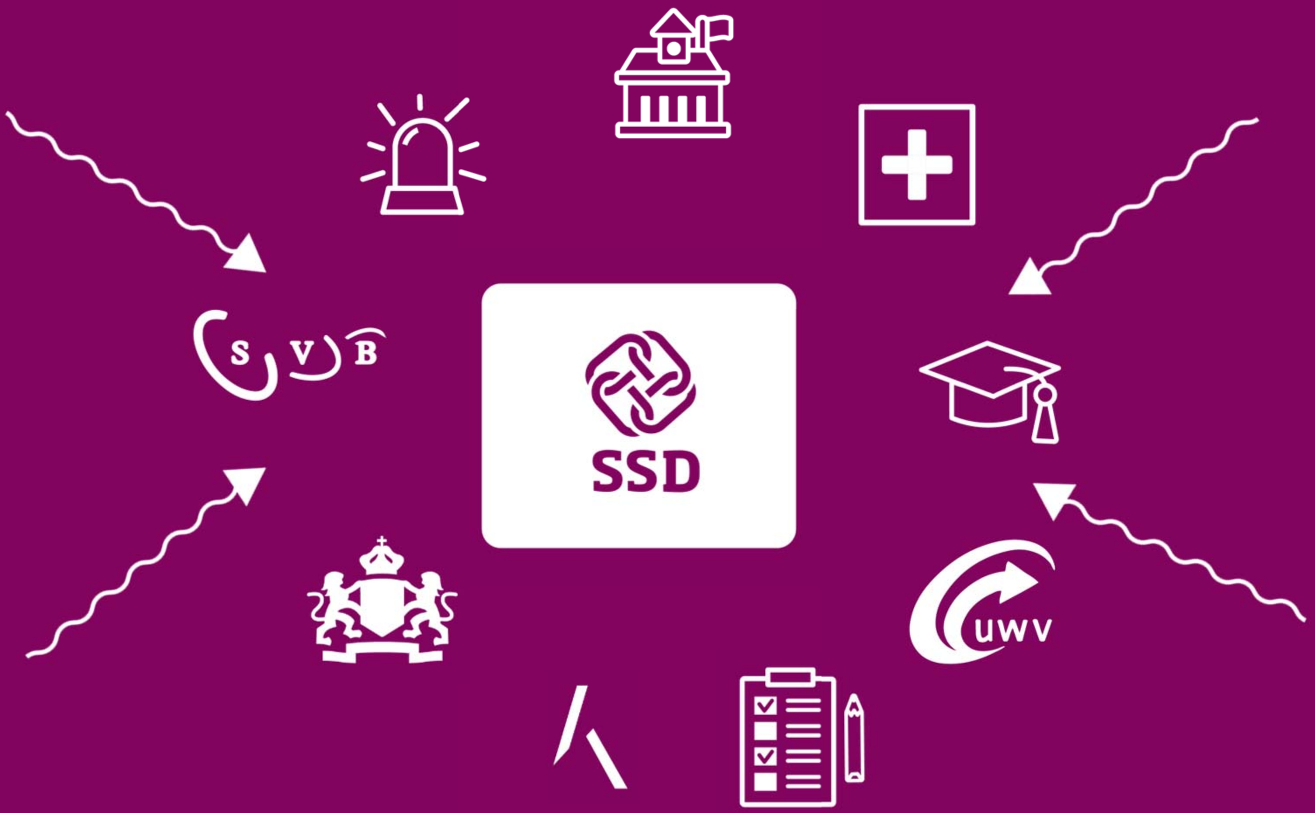
1. Registers (i.e.): Citizen Service Number (unique), addresses
2. Surveys (i.e.): Sex, date of birth, address (postal code and house number)

Matching	Return
Citizen Service Number	>95%
Postal code6+sex+date of birth	>90%
Postal code4+sex+date of birth	>85%

Joining: the SSD

SSD
School
Office

Data from different sources



The reason

SSD

Changing society

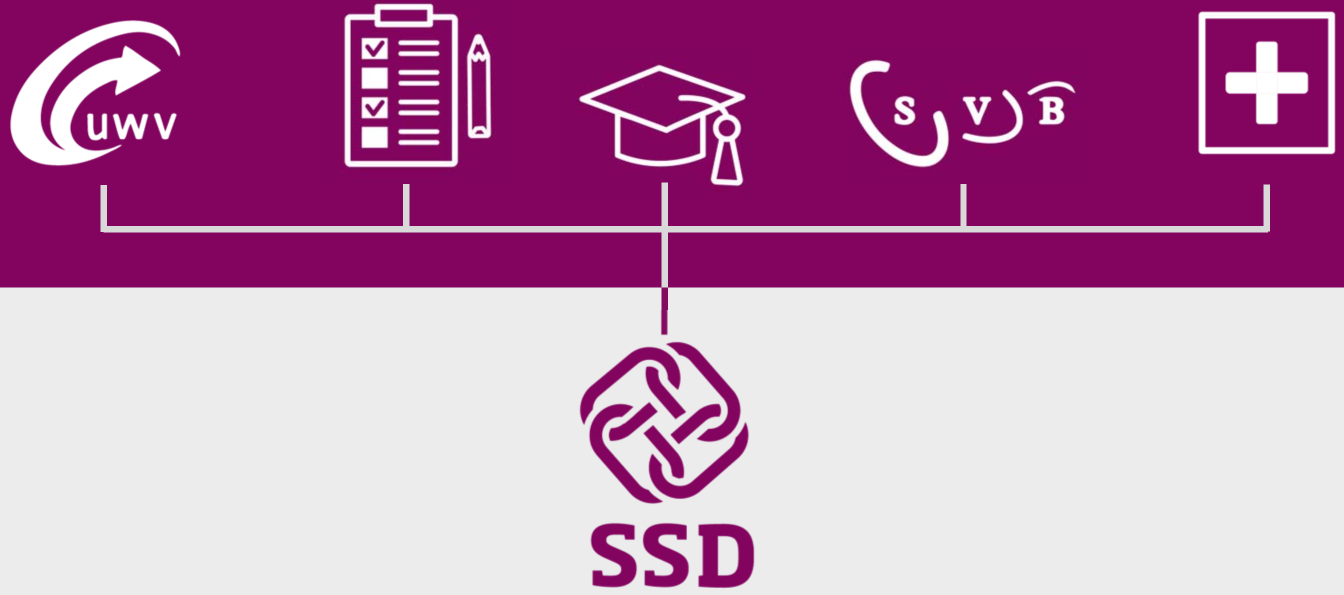
More possibilities

Questions of today

Then:

Now:

The SSD structures and links data



The strength of the SSD

Long period

Combine

Specific

High quality

www.CBS.nl

@Statistiekcbs

E.Gerards@cbs.nl

Egon Gerards
Head of SSD department

PRIVACY audit proof

Privacy

Checked for disclosure risks

Anonymised data

Limited access

Only output is disseminated, never any data

How does the SSD work?