Synthetic Data

A tool to enhance open science practices

Katharina Krüsselmann & Jim Achterberg

Universiteit Leiden
The Netherlands

Marcel Haas (Leiden University Medical Center)

Marco Spruit (Leiden Institute of Advanced Computer Science)

Marieke Liem (Institute of Security and Global Affairs)

Project SENSYN

Finding alternative ways to make data FAIR

Guidebook for the use of synthetic data

- How-to's: which methods fit which type of data & purpose
- Advantages & challenges
- accessible

Proof-of-concept

- Interactive platform (targeted at non-academic audiences) with synthetic data based on Dutch Homicide Monitor

Open Science Fund

The Open Science Fund aims to support projects specifically designed to implement and stimulate open science practices. With this funding instrument, NWO takes a step forward towards changing the way academics are recognised and rewarded in the Netherlands.

Purpose and objectives

NWO wants to stimulate open science by incentivizing and rewarding researchers from all disciplines who are or would like to be at the forefront of this movement.

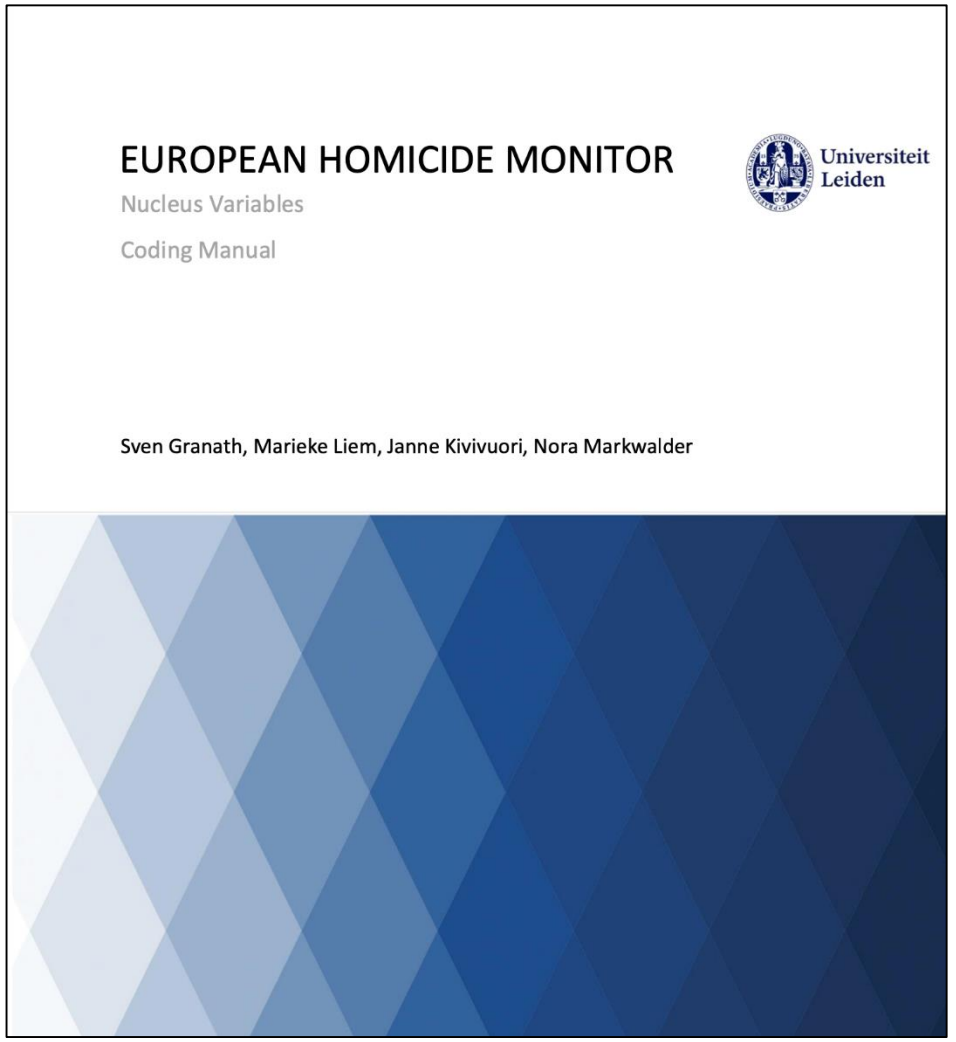
Particularly encouraged are projects that:

- improve how good open science practice is recognised and rewarded;
- transform the way researchers publish;
- develop or adapt interoperability standards;
- develop, test or adapt open platforms or tools for wider community use;
- stimulate wider adoption of open science practices among researchers;
- further the adoption of citizen science approaches

Dutch Homicide Monitor

Individual level data of homicide victims and perpetrators

Personal data: age, gender, profession, criminal history …

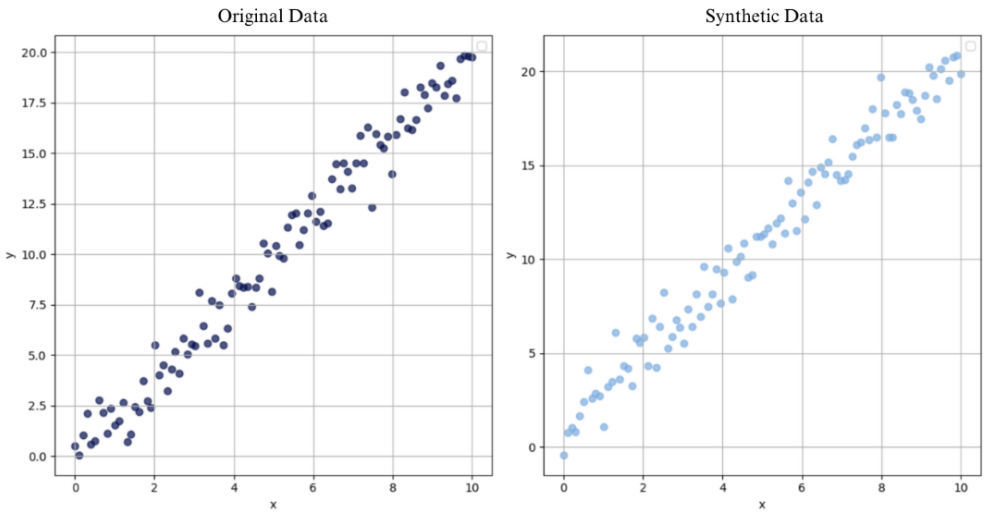


Synthetic Data

Artifically manufactured data

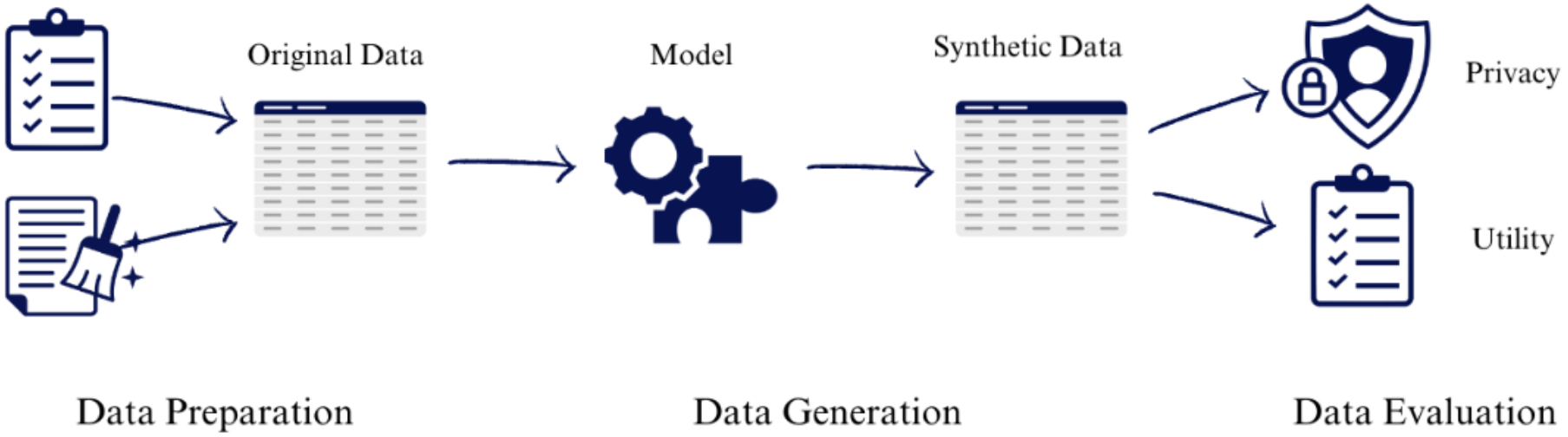
Structured data (quantitative) & unstructured data (text, visual)

Digital Twin > characteristics & specific parameters of original data are similar



(Krüsselmann et al., 2024)

Generating Synthetic Data



(Krüsselmann et al., 2024)

Promises of Synthetic Data

Privacy preservation

Data & environment manipulation

Cost & time efficient

Enables data accessibility & reuse

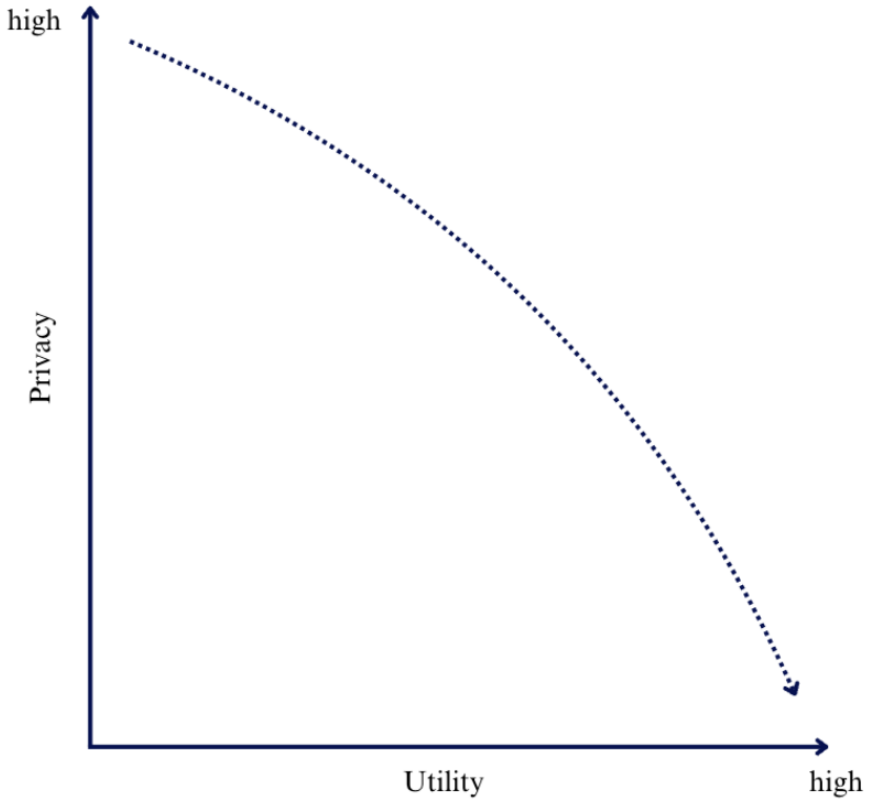
Risks of Synthetic Data

Quality & privacy **not guaranteed per se**

Garbage in, (worse) garbage out

Complexity of data generation

Lack of standards



Synthesis of the Dutch Homicide Monitor

Categorical data

21 variables

Complex structure

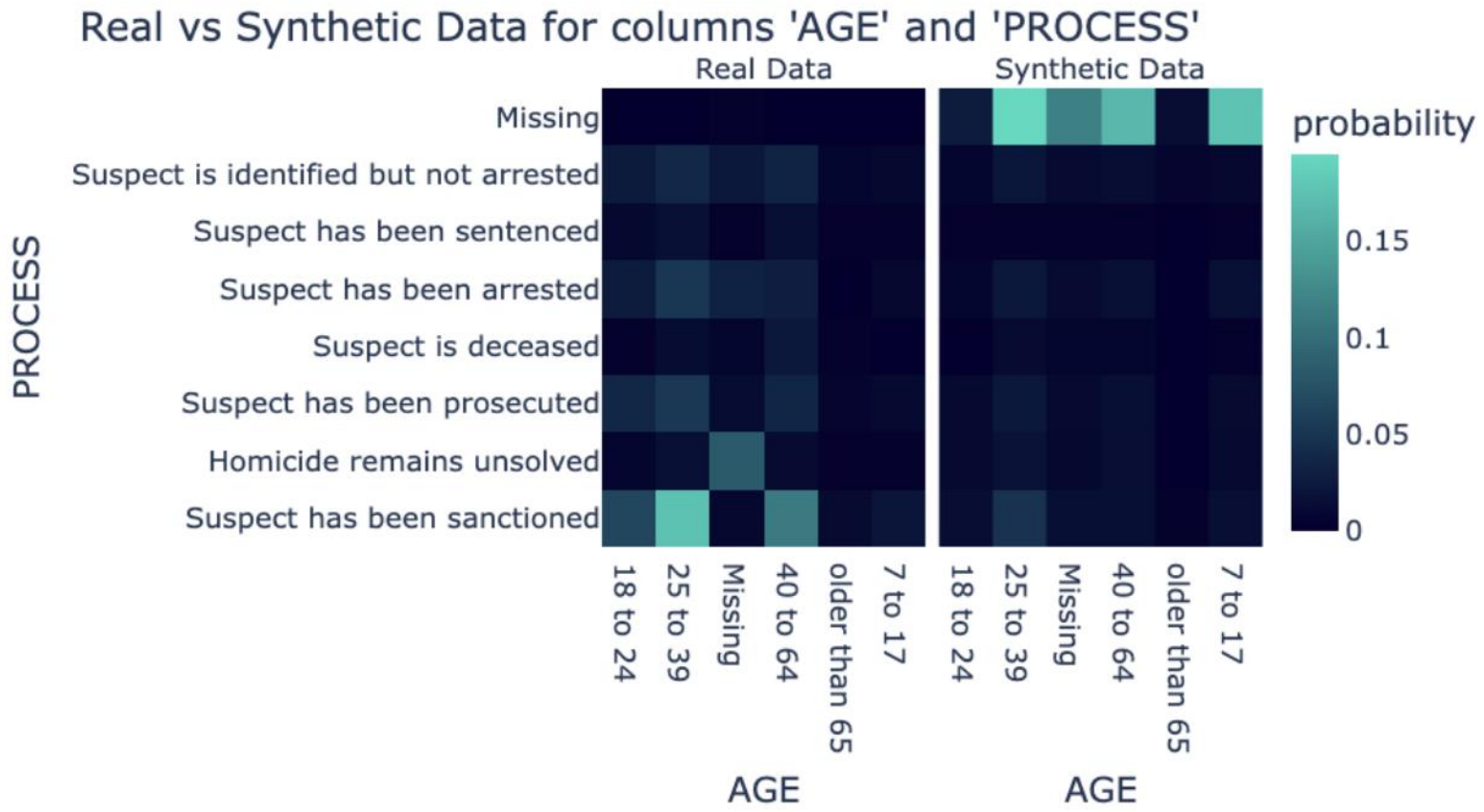
		Case-level		Individual-level		
Case 1	Victim	Bar	Morning	Male	34	Intoxicated
	Perpetrator	Bar	Morning	Male	48	-
Case 2	Victim	Home	Evening	Female	21	Missing
	Perpetrator	Home	Evening	Male	57	Drug

Goal synthetisation

- Capture full structure and detail
- Same/similar univariate & bivariate trends
- Privacy safeguarded

Synthesis of the Dutch Homicide Monitor

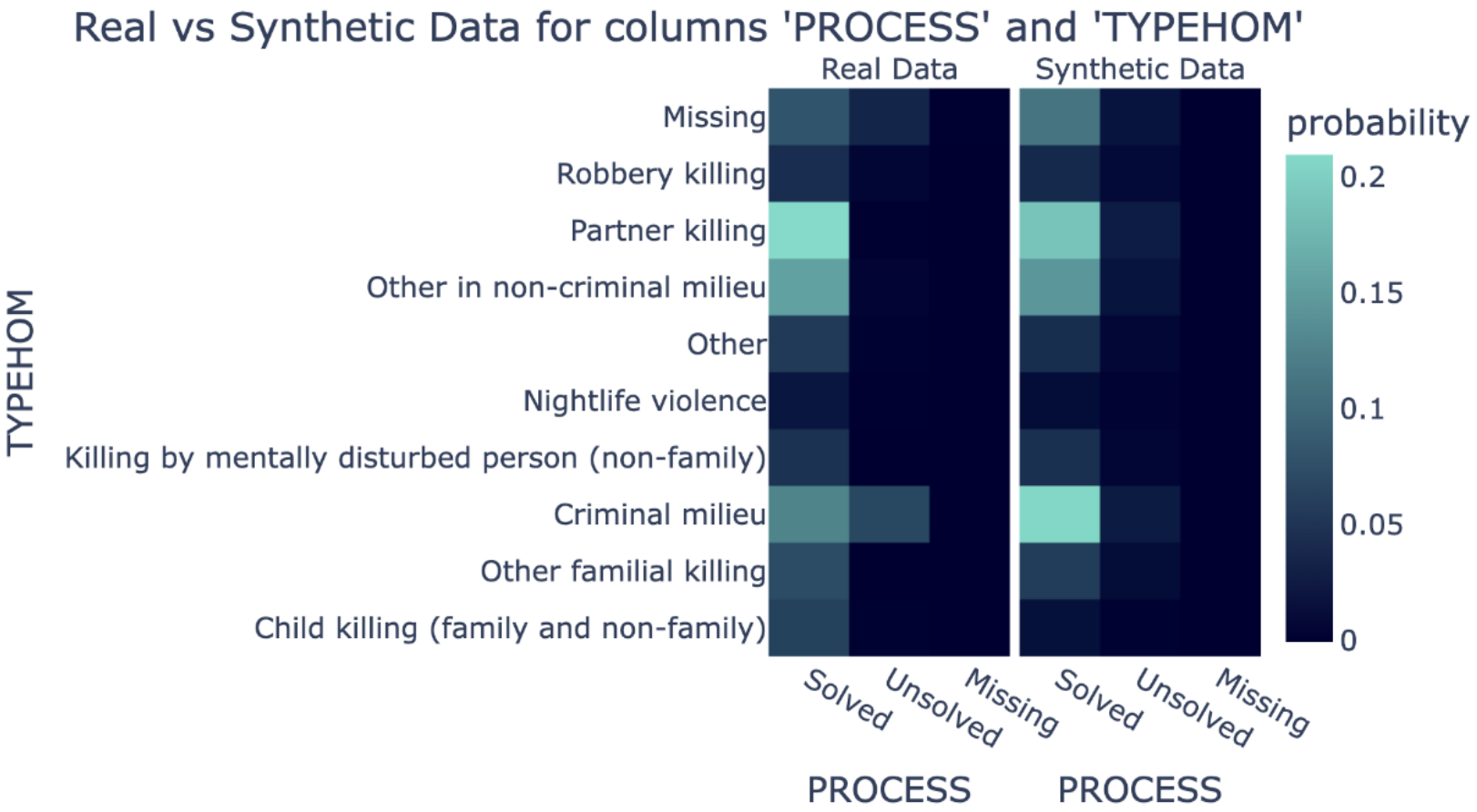
First attempt



Synthesis of the Dutch Homicide Monitor

Final attempt

Quality Type	Score (%)	In
Data validity	100	Th as
Data structure	100	Th wi
Univariate distribution	97.8	A 98
Bivariate distribution within one table	89.67	Th sy re
Overall	93.74	O ar



Project SENSYN

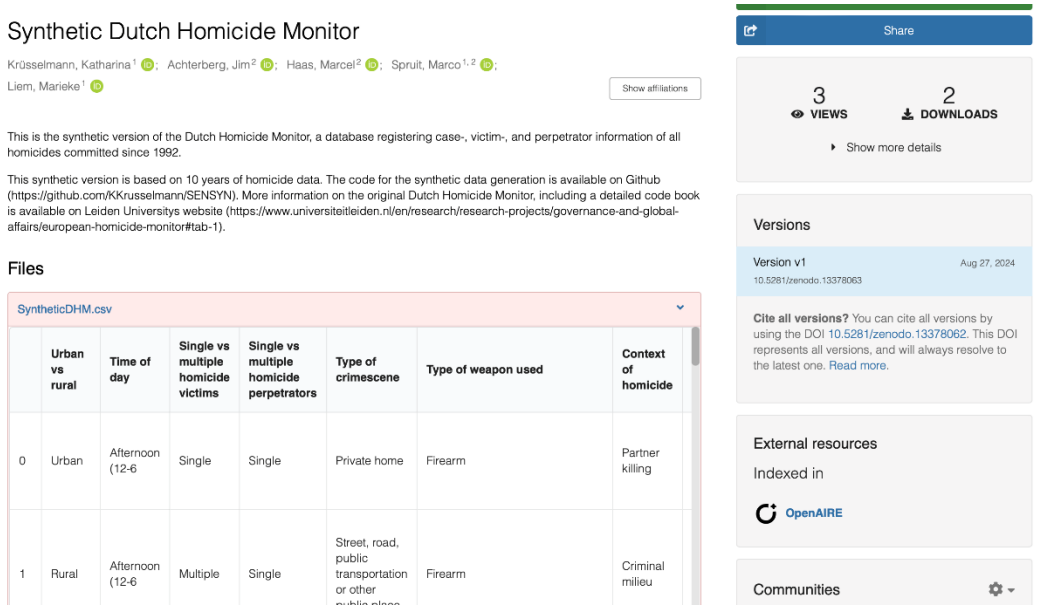
Guidebook (open access)

Open code (github: KKrusselmann/SENSYN)

Open synthetic dataset (Zenodo repository with DOI)

Web application on synthetic data, including synthetic dataset

dutchhomicide.streamlit.app



Synthetic Dutch Homicide Monitor

Krüsselmann, Katharina[1]; Achterberg, Jim[2]; Haas, Marcel[2]; Spruit, Marco[1,2]; Liem, Marieke[1]

Show affiliations

This is the synthetic version of the Dutch Homicide Monitor, a database registering case-, victim-, and perpetrator information of all homicides committed since 1992.

This synthetic version is based on 10 years of homicide data. The code for the synthetic data generation is available on Github (https://github.com/KKrusselmann/SENSYN). More information on the original Dutch Homicide Monitor, including a detailed code book is available on Leiden Universitys website (https://www.universiteitleiden.nl/en/research/research-projects/governance-and-global-affairs/european-homicide-monitor#tab-1).

Files

SyntheticDHM.csv

	Urban vs rural	Time of day	Single vs multiple homicide victims	Single vs multiple homicide perpetrators	Type of crimescene	Type of weapon used	Context of homicide
0	Urban	Afternoon (12-6	Single	Single	Private home	Firearm	Partner killing
1	Rural	Afternoon (12-6	Multiple	Single	Street, road, public transportation or other	Firearm	Criminal milieu

Share

3 VIEWS 2 DOWNLOADS

Show more details

Versions

Version v1 Aug 27, 2024

10.5281/zenodo.13378063

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.13378062. This DOI represents all versions, and will always resolve to the latest one. Read more.

External resources

Indexed in

OpenAIRE

Communities

Thank you

Universiteit
Leiden
The Netherlands

Discover the world at Leiden University