An Introduction

Tom Emery, Executive Director
A Social Science Infrastructure

The ODISSEI council has 27 members that together contribute €1 million per year. The number of members is continuously growing and diversifying.
A Social Science Infrastructure
Main Activities of CDISSEI

**Microdata Services**
Secure access to CBS data for around 700 researchers from a broad range of scientific disciplines.

**LISS Panel**
A national laboratory for high quality methodological and scientific innovations.

**Financing Fieldwork**
Maintaining existing panels such as SHARE, ESS, EVS and GGP. Promoting open access data collections in the social sciences with a good price-quality ratio.

**HPC Access**
Development of a secure HPC environment for social scientists for the analysis of survey and register data.

**Standards**
Definition and implementation of standards within the social sciences to increase interoperability, monitoring change and improving data quality.

**Communications**
Building bridges with other disciplines, representing the interests of social scientists and participating in international initiatives.
ODISSEI Goal: A World First

An infrastructure that brings together register data, survey data and high performance computing in a digital environment that is available to the entire community of social scientists.
CBS Microdata

- Individual Population Register
- Organizational Registers (Companies, Charities etc)
- Tax Records

- High quality Persistent Identifiers everywhere
- Pseudo Anonymized data is available for research
CBS Microdata

Microdata

On Site

Analysis

Remote Access

Upload

Output

Analysis

Output
ODISSEI Data Facility
What can we do?

Use Cases in the ODF
Geo-Spatial Analysis

- Precise mapping of specific populations or variables
- For example, % of population with ‘Non-Western’ heritage
- Calculated at small scale 100m x 100m (n = 308,010)
- Calculation repeated for increasing catchment area to measure entropy
A) 100m×100m grid cells
C) 2000m radius

Share of people with a non-Western background (%):

- 0-10
- 11-25
- 27-45
- 47-69
- 70-100

Legend:
- Municipality of Amsterdam
- Water
- Transportation site
- Built environment
- Green space

Source: Calculations based on data from the Statistics Netherlands (CBS)
This took four months to run at CBS on one Node

ODF allows us to distribute the job

This allows analysis to expand to temporal analysis multiple indicators

Resulting data can also be integrated into any other ODF analysis

Genetic Analysis

- National Twin Registry (11,000 Genetic Samples)
- Linking Genetic Data to Registers (with Consent!)
- Data is big, computation is demanding (GWAS)
- One person's genetic data is bigger than the register

- Too big for CBS
Genetic Analysis
Genetic Analysis

Median genetic risk score for schizophrenia in Dutch Towns
Genetic Analysis

- Genetic Analysis is already done at SurfSara
- The ODF can allow the enrichment of medical data with Administrative data
- Better environment for interdisciplinary research

Social Network Analysis

- CBS data used for demographic research to match partners or parents and children etc.
- But...the main constraint is computational
- Small scale sub sample analysis: 4,000 links per person
- Scaling up would be 72 billion links for the whole Dutch population
Social Network Analysis

Family Networks
Social Network Analysis

Colleague Networks
Social Network Analysis

School Networks
Many, many possibilities

Network Linkages made can be reused in the ODF

E.g. calculating entropy of specific characteristics across social network and geospatial dimensions simultaneously
Next Steps

- It works but…
- ETHICS
- Data management
- Infrastructure Sharing
- Skills Gap between computational and social sciences
- Rapid technological developments
Next Steps

- Large Research Infrastructure Proposal
- Scaling Up the ODF
- Improving the interface
- Creating a Community
- Integrating data collections
- Innovating in data collections
THANK YOU