Value added data archive service in using admin data for teaching and research in Slovenia

Janez Štebe, ADP, UL, Ljubljana

Access to anonymised administrative data for the purposes of research and policy analysis – where are we 10 years after the adoption of the Hungarian Act 101 of 2007?

A workshop for the Visegrad countries and beyond

21-22 September 2017

Venue: Hotel Benczúr, 1068 Budapest, Benczúr u. 35.

ADP ARHIV DRUŽBOSLOVNIH PODATKOV SOCIAL SCIENCE DATA ARCHIVES

Within its mission, the ADP establishes itself as a national infrastructure that:

- ingest important data sources from a wide range of social sciences, interesting for analyzing the Slovenian society,
- preserves
- and promotes their further use in scientific, educational and other purposes.

http://www.adp.fdv.uni-lj.si/

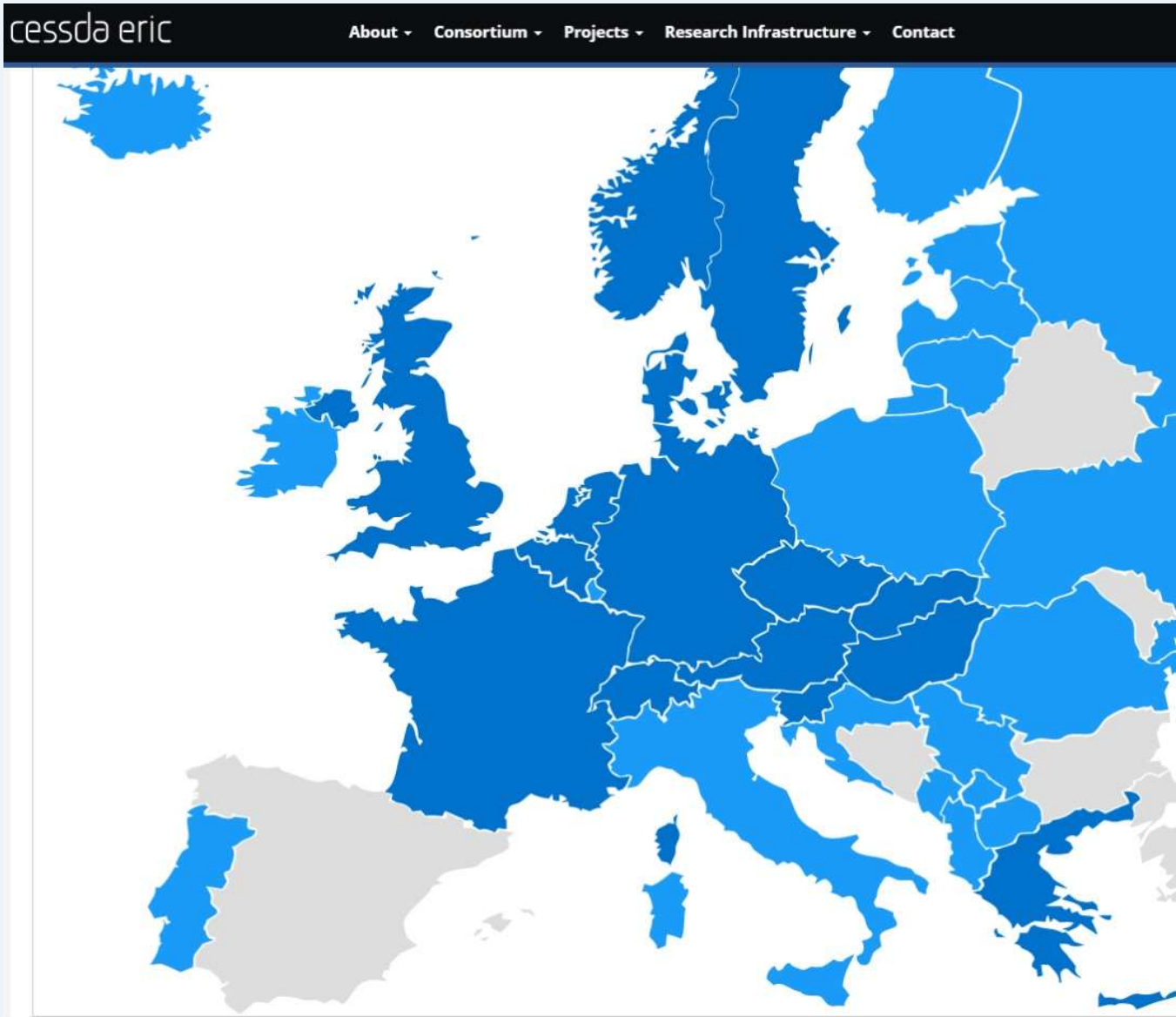
Administrative data

- Administrative data are defined as data which derive from the operation of administrative systems, typically by public sector agencies. They cover activities such as health maintenance, tax and social security, housing, elderly care, vehicle and other licensing systems, educational progress, etc. While such data are not designed for research purposes, they often have significant research value, especially when linked to other datasets or to user-generated surveys.
- Peter Elias: Administrative Data. In **Facing the Future : European Research Infrastructures for the Humanities and Social Sciences. http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:kobv:b4-opus-26346**
- See also **Administrative Data Liaison Service** at http://www.adls.ac.uk/adls-resources/guidance/introduction/

CESSDA stands for Consortium of European Social Science data Archives and ERIC stands for European Research Infrastructure. CESSDA provides large scale, integrated and sustainable data services to the social sciences. It brings together social science data archives across Europe, with the aim of promoting the results of social science research and supporting national and international research and cooperation.

CESSDA ERIC members

CESSDA partners



Access to public administrative data in Slovenia for research

- National Statistical Institutes of Slovenia (SURS) obtains the admin data from government institutions for the reporting purposes
  - Admin data sources that are declared in the **Programmes of statistical surveys (**http://www.stat.si/statweb/LegislationAndDocuments/StatSurveys)
    - e.g. Agency of the Republic of Slovenia for Public Legal Records and Related Services (AJPES) provides data from Business Registers
- SURS can provide access to those admin data further for research purposes

Access to official statistical micro data at the Statistical Office of the Republic of Slovenia

- **Modes of access and type of data;**
  - On-site laboratory: ScUF (Secure Use Files)
    - primary protected (standard identifiers are removed)
    - researchers link and analyse data
    - final output is checked against confidentiality by SURS
    - free of charge
  - Remote access facility: ScUF
    - researchers link and analyse datasets
    - final output is checked against confidentiality by SURS
    - charges apply
  - CD: SUF (Scientific Use Files)
    - researchers can get limited number of social surveys with less than 1% population included
    - primary and secondary protected data prior to analysis
    - free of charge
- **Users:** licenced researchers; has to apply and justify the request; sign contract and confidentiality declaration
- **SURS;** confidentiality committee assess the request, prepares the proposal for board of directors, director general decide

An overview of cooperation between ADP and SURS

- has been going on since the establishment of the Slovenian Social Science Data Archives (ADP);

1. ADP has been distributing added value PUF (Public Use Files): Anonymised SORS microdata (selection of variables, samples) and structured metadata
   - variable and value labels, missing values are added to the dataset; additional logical control is made, unneeded variables are deleted, variables in databases are connected to codebooks used
   - Accessible for registered users, including students, at ADP (no additional restrictions)
2. For data otherwise accessible in the Safe room at SURS:
   - ADP provided enhanced documentation and comprehensive metadata, available and searchable on the catalogue on internet
3. Promotion activities for microdata use, for research and education purpuses
4. Both organisations were partners in DwB (Data without Boundaries) project, together with many other CESSDA and Official statistics organisations

List of microdata from official statistics that are listed in ADP Catalogue (Bold=PUF available)

Study ID	Study Title	List of Studies
ADS	**Labour Force Survey**	ADS, ADS14, ADS13, ADS12, ADS11, ADS10, **ADS10P**, ADS09, ADS08, ADS07, ADS06, ADS05, ADS04, ADS03, ADS02, ADS01, ADS004, ADS003, ADS002, ADS001, ADS00, ADS994, ADS993, ADS992, ADS991, ADS984, ADS983, ADS982, ADS981, ADS974, ADS973, ADS972
MBSSMM	**Microdataset of Social Statistics for Development of Microsimulation Model**	MBSSMM10, MBSSMM07
POPIS	**Register Census**	POPIS, POPIS11, **POPIS11p**, **POPIS02**
APG	**Household Budget Survey**	**APG00**
APC	**Time Use Survey**	**APC01**
AZK	**Crime Victim Survey**	**AZK01**

Register Census 2011: Data with no confidentiality restrictions

Teaching with real official microdata

- Based on experience of promotion of LFS PUF data in classroom, SURS and ADP agreed to initiate CUF (Campus Use File) preparation and promotion

- Characteristics:
  - Collaboration among substantive topic expert, methodology expert from SURS and ADP staff

- Exploratory workshop was organised in September 2016, and meetings organised in the following months:
  - Recrutation of interested professors (substantive contribution expected and in return, individual costume made data file for classroom use was promised by SURS staff).

Approach

- It was agreed that teaching resources pack will consist of:
  - CUF (Campus Use File), that methodology and statistical confidentiality expert from SURS will prepare
  - Enhanced documentation and metadata will be prepared by ADP staff, data will be accessible through ADP under the CC0 licence
  - Accompanying guide covering data access, statistical analysis and introduction to substantive topic and concepts, with research questions and examples of exercises and solutions.
- Three areas of interest for which separate data files will be constructed
  - Family
  - Migrations
  - Youth

CUF Family: design process

- Professors Alenka Švab and Andreja Živoder from UL suggested the topics and variables, Danilo Dolenc from SURS advised
- Mid population census microdata from 2015 served as a source
- The original data is produced by linking several administrative registers
- Manca Golmajer from SURS Anonymization department prepared final CUF
- Limitations:
  - 5% sample of individuals, members of families stratified based on Sex, Age and Family type
  - For practical reasons of anonymizations:
    - Ten variables maximum
    - Reduced number of categories for each variable

CUF file quality characteristics

- Confidentiality restrictions
  - N: 83.944.
  - N (proportion) of units that doesn‘t contain any statistically protected data: 81.920 (97,6 %).
  - Each variable has less then 1 % statistically protected data.

Variable name	**Meaning**
ID	Id
TIP_DRUZ	Family type
GO_POL1	Person family role
VZP_DRUZ_IND1	Indicator of Reconstituted family
STAR1	Age
SP	Sex
ST_OT_DRUZ1	N of children living in family
ROJ_ST1	Children born
AKT2	Activity status
IZB2	Education
ST_URB1	Urbanisation
UTEZ	Weight

Remaining tasks and conclusions

- Observations: Both supply of data for research and teaching and demand for such data is a problem
- ADP role to find a common denominator between demand and supply:
  - CUF project is ideal with respect not to put to much burden on SURS data protection unit; supply granted
  - reduced complexity of data, and users‘ guidance provided which introduces basic data analytic skills, makes data accessible to relatively statistically illiterate audiences (including the lecturing staff); demand and use granted
- Next generation of advanced administrative microdata users will grow