Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
Consiglio Nazionale delle Ricerche

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

FOSSR ONLINE SHORT COURSES

MODULE 1 Data life cycle, data curation and preservation

October 4-5, 2023

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES: Agenda October 4, 2023

Introduction: Andrea Orazio Spinello (CNR- IRCrES , Leader of the FOSSR project WP on training activities)

Module 1. Data life cycle, data curation and preservation
Data lifecycle. The module will address the main issues that can arise from design to production and storage and finally to access and analysis. It therefore offers the opportunity to analyze and reflect on the entire production plan on an important occasion, especially for those approaching these issues for the first time. The different types of data, typical of the social sciences, must necessarily be treated in different ways and methods that are appropriate to the data's measurement level. Students will have the opportunity to create sample databases designed to apply a reading of information with statistical methods.
Organization and teaching: Loredana Cerbara, Nicolò Marchesini

Module 2. Data curation and conservation
The main methods and tools for privacy and compliance with the GDPR 2016 are illustrated. The approach chosen in the CNR area will be used for managing personal data, risk analysis, implementation of actions and tools for risk reduction to guarantee the rights of data subjects.
Organization and teaching: Loredana Cerbara

ONLINE LABORATORY

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES: Agenda October 5, 2023

Module 3. Data typologies and architectures in the social sciences

Some social science platforms for open data sharing (eg CESSDA - Consortium of European Social Science Data Archives - or Zenodo , RISIS, SHARE, GGP but also Italian Online Probability Panel (IOPP)) will be shown, in line with open science principles. Particularly, some platforms (such as DASSI - Social Sciences Data Archives Italy -, Italian node of CESSDA) will be shown as examples of data production and processing and as a possibility interoperability between platforms.

Organization : Loredana Cerbara, Nicolò Marchesini – Teaching: Filippo Accordino , Loredana Cerbara, Gabriella D'Ambrosio , Nicolò Marchesini, Luciana Taddei , Emanuela Varinetti and Francesco Visconti

Module 4. Statistics for the social sciences

A general overview of the approaches and families of methods generally used is necessary, which can constitute a guide for those approaching this topic. For this reason an overview of quantitative and qualitative analysis will be done and students will directly experience the application of some quantity methods with sample data

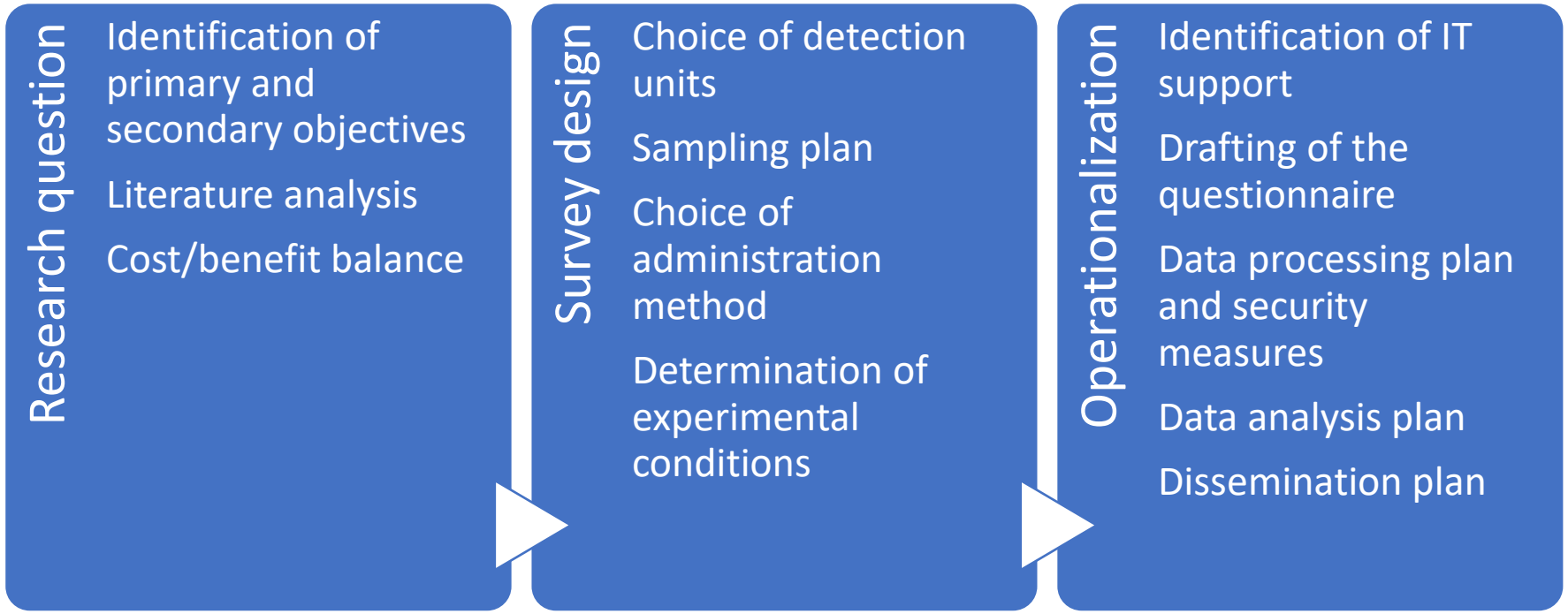
Organization and teaching: Loredana Cerbara, Nicolò Marchesini

ONLINE LABORATORY

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

Module 1. Data life cycle, data curation and preservation

1-DESIGN



DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

Module 1. Data life cycle, data curation and preservation

2-CREATION

Field – field investigation

- Selection of detection units
- Administration
- On-field control and possible introduction of corrective elements

Data cleaning and validation

- Identification and treatment of outliers
- Identification and processing of missing data
- Creating recoded variables that can be used for analysis
- Creation of synthetic indexes

Analysis and reporting

- monovariate and multivariate statistical analyses
- Reading and interpretation of results
- Reporting
- Disclosure

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

Module 1. Data life cycle, data curation and preservation

1-DESIGN

Research question : _

- Identification of primary and secondary objectives, i.e. answering the questions:
  - What is my search for?
  - What areas does it have an impact on?
  - Who is it useful for? Who might be interested in ?
- Analysis of the literature in search of similar experiences, to understand methods and methodologies, strengths and weaknesses
- Cost/benefit balance. Given a research effort that has a real cost, do the benefits have an equivalent or greater (even potential) value?

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

Module 1. Data life cycle, data curation and preservation

1-DESIGN

Drawing of the investigation :

- Choice of survey units, i.e. to answer my research question on which survey unit should I base the design?
- Sampling plan, i.e. the set of decisions that lead to the survey unit starting from a ***reference universe that we must study and understand because all subsequent decisions depend on it*** . It must be a type of sampling suitable for the purposes of the research but under the constraint of economic availability (both strictly economic and available personnel). It can be *probabilistic* or *non-probability* sampling , *with one or more stages* , *by quota or proportional* or without proportionality constraints. (next SLIDE focus)
- Choice of administration method (CATI, CAPI, CAWI, PAPI, ...)
- Determination of the **experimental conditions** , i.e. the rules that allow detection in the same conditions for all statistical units.

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

1-DESIGN: FOCUS ON THE SAMPLING PLAN

PROBABILISTIC SAMPLES (inference allowed)

- Availability of a list of statistical units (detailed knowledge of the Universe)
  - *simple* or *stratified* random sampling (+), *single or multi-stage* (+), simple or *cluster sampling* (-). The extraction can also take place *in bulk* (-) or in *systematic mode* (-) with random choice of the first unit (foot of the distribution with random choice)
  - The sampling plan must always keep *the probability of inclusion* of statistical units constant and this probability must not be zero for any unit of the Universe (all units must be able to be selected for the sample) under penalty of forfeiture of the randomness or validity of the sampling facility.
- NON-PROBABILISTIC SAMPLES (inference not allowed)
  - A list of statistical units is not available. The probability of inclusion cannot be calculated, some units may never be considered. This methodology more often causes bias in estimates , as units included in the sample may have different characteristics than those not included.
    - It is important to collect as much information as possible about the Universe.
    - Examples *: reasoned choice sampling , quota sampling, avalanche sampling* .

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

1-DESIGN: FOCUS ON THE SAMPLING PLAN How many cases?

The most frequent question is: how many interviews do I need to do to make the sample reliable?

- Answers based on mathematical methods:
  - You must consider the desired precision (reliability of the estimate), the width of the confidence interval (range) and the type of estimate you want to obtain. For estimating a proportion (e.g. response rates) for a 95% interval (the probability that the estimate falls within the chosen interval is 95%) with a precision of D, we obtain: $n = \frac{z^2\, p(1-p)}{D^2}$ where ***n*** is the number sought, ***z*** is a value obtained from the probability distribution associated with the estimate, ***p*** is the expected proportion, ***D*** the desired level of precision for the estimate. If we approximate (often we can do so) the probability distribution with the Normal z is equal to 1.96, p can be set equal to 0.5 because the multiplication by 1-p in this way takes on a maximum value, D is of choice of the researcher.
  - EXAMPLE: we want to know how large a sample must be to estimate the percentage of respondents in favor of an initiative with an accuracy of 5%. $n = \frac{1{,}96^2\ 0{,}5(1-0{,}5)}{0{,}05^2} = \frac{3{,}8416*0.25}{0{,}0025} = \frac{0{,}9604}{0{,}0025} = 384{,}16$

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

1-DESIGN: FOCUS ON THE SAMPLING PLAN How many cases?

The most frequent question is: how many interviews do I need to do to make the sample reliable?

- We note that:
  - The result is based only on the precision of the estimates and does not depend on the size of the reference Universe.
  - The choice of the reference statistic depends on the type of estimates you intend to make (for example, it could be an arithmetic mean or a complex index)
  - If the variability of the estimates in the Universe is known, for example because a previous study is available, it can be taken into account when determining the sample size.
- Answers based on empirical methods :
  - Economic and personnel resources must be taken into account (cost constraints)
  - The time window must be taken into account (time constraints)
  - It is necessary to study the minimum characteristic that we want to be representative to answer the research question (representativeness constraints): for example, if for a survey on families it is necessary to study a phenomenon in particular with respect to the presence or absence of young children with respect to the territorial distribution or to the socio-economic status, I need to have a sufficient number (generally above 100 units) in the target stratum, with consequent proportionation of all the other strata

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

Module 1. Data life cycle, data curation and preservation

1-DESIGN

Operationalization :

- Identification of the IT support, i.e. the software and hardware that will support the detection
- Drafting of the questionnaire, identification of the ***structural variables*** of interest for the data analysis and of the specific information questions for answering the research questions. We dedicate a laboratory activity to this .
- Data processing plan and security measures: we dedicate a separate slot to this
- Data analysis plan. Predict which analyzes will be carried out and try to formulate the questions so that they can then be used in the analyzes you want to conduct. For example, open-ended questions generally do not allow multivariate analyzes unless after specific and dedicated processing.
- Dissemination plan

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

Module 1. Data life cycle, data curation and preservation

2-CREATION

Field – investigation in the field:

- Selection of detection units. The decisions made during the sampling plan phase are applied and the list of units to be interviewed is arrived at starting from the reference Universe
- Administration. Depending on the administration method chosen, the statistical units are interviewed.
- On-field control and possible introduction of corrective elements. During the field it must always be checked that the survey proceeds as planned and distorting elements that were not originally foreseen must be identified as soon as possible. What is established by the experimental conditions must always be maintained to prevent the administration from being different for the different statistical units, introducing systematic errors that distort the estimates (for example interference during administration by other individuals not involved, such as teachers when administering questionnaires to students or excessive comparison between students who all give the same answers, etc.)

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

Module 1. Data life cycle, data curation and preservation

2-CREATION

Cleaning and validation of the data (to this we dedicate a activity laboratory)

- Identification and treatment of outliers , i.e. data that are very different from the rest of the distribution which can be generated due to insertion errors or even because they are rare events that fall into the sample. They can be treated in many ways (elimination, recoding of variables with extremes in the same class, specific treatment, etc.)
- Identification and processing of missing data, both through automatic analyzes that can replace missing data with similar values detected on units of the same type, and with *congruity analyzes* (for example, Italian citizenship can be assigned to pupils who have both Italian parents, one age group estimated for pupils of the same class, etc.)
- Creation of recoded variables that can be used for analysis, i.e. starting from one variable it is possible to generate multiple variables useful for subsequent analyzes (e.g. reduction into classes of different size)
- Creation of synthetic indexes

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

Module 1. Data life cycle, data curation and preservation

2-CREATION

Analysis and reporting

- monovariate and multivariate statistical analyses . An initial monovariate analysis is always useful , i.e. the frequencies of the answers to each question, or the representative values if they are continuous variables, to verify the validity of the previous operations. Bi- and multivariate analyses, on the other hand, are destined to produce information that is not evident at first but is always the most interesting. We dedicate a slot to this .
- Reading and interpretation of results. We dedicate a laboratory activity to this
- Reporting
- Dissemination (organisation of study days and conferences, press releases, populating websites and social media, etc.)

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
Consiglio Nazionale delle Ricerche

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

FOSSR ONLINE SHORT COURSES

MODULE 2. Data curation and conservation

October 4-5, 2023

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

DATA LIFECYCLE AND ARCHITECTURES IN THE SOCIAL SCIENCES

Module 2. Data curation and conservation

MATERIALS AVAILABLE ON THE CNR DPO (OR RPD) WEBSITE www.rpd.cnr.it

GET YOUR ORIENTATION

- **Reference legislation**
  - The General Data Protection Regulation
  - The Privacy Code
  - The Ethics Rules
  - The provisions of the Guarantor for the Protection of Personal Data
  - The Declaration of Helsinki
- **Main concepts**
  - The data controller
  - Joint ownership in a treatment
  - The data processing manager
  - The principles of the GDPR
  - The conditions of the processing (the legal bases)
- **Internal organisation**
  - Processing of personal data at the CNR
  - The Data Protection Officer
  - The correspondents of the Data Protection Officer
  - The Working Group supporting the General Director
  - The Privacy Contacts