USING EXISTING DATA

Secondary data analysis

What is secondary analysis?

Primary data is data we collect ourselves and

Secondary data is that collected by others

Secondary analysis is done on secondary data

In other words, someone else gathered the data – for their own purposes – and then we analyse it for our own purposes.

General observations

- A large proportion of research is based on secondary data
- The issues encountered in using secondary data are similar to data issues in other context
- There is a need for a research community for the sharing of secondary data;
  - *Making data available in the public domain*
  - *Data evaluation and quality check*
- New information from the same data, because of new analytical tools, new theoretical perspectives, and new operationalization
- The possibility of further use (reanalysis of data)

Issues related to the use of secondary data

- An observation
  - *issues are similar to data issues in other types of empirical research*

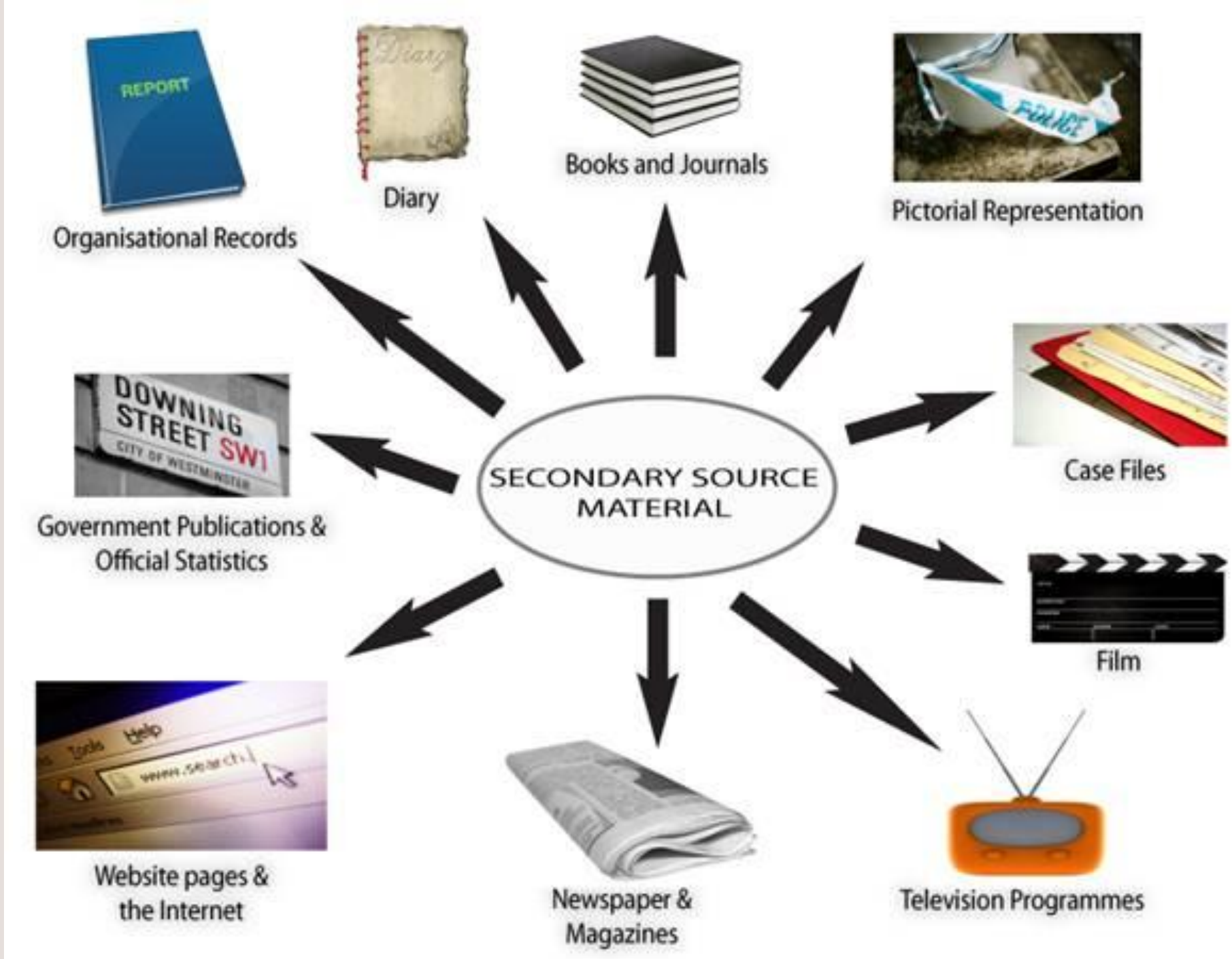
- Assessment of data quality
  - *The purpose, information of the data*
  - *The population of study, sampling framework and procedures*
  - *Methods of data collection, response rate*
  - *Data coding and entry*
  - *Codebook – questionnaire, coding scheme, etc.*
  - *Previous research using the data*

Advantages of secondary analysis

- Saves money and time
- Offers high quality data
- Gives an opportunity for longitudinal analysis
- Allows subgroup or subset analysis
- Gives an opportunity for cross-cultural studies
- Allows more time for data analysis
- Enables the application of recent theory to old data
- Gets more value from the original data

…but there is a down-side…

- You need to become familiar with how the data was collected, coded and managed
- The data can be very large and complex
- The quality of the data should never be taken for granted
- Variables important to your analysis might be missing


REPORT
Organisational Records
Diary
Books and Journals
Pictorial Representation
DOWNING STREET SW1
CITY OF WESTMINSTER
Government Publications & Official Statistics
SECONDARY SOURCE MATERIAL
Case Files
Film
Website pages & the Internet
Newspaper & Magazines
Television Programmes

Examples of large data sets suitable for secondary analysis

British Household Panel Survey (BHPS); now Understanding Society – The UK Household Longitudinal Study	A panel study that began as the BHPS in 1991 and was conducted annually by interview and questionnaire with a national representative sample of around 10,000 individuals in just over 5,000 households. The same individuals were interviewed each year. The BHPS was replaced in 2010–11 by the Understanding Society survey, which is based on a much larger panel of 40,000 households and which incoprorates the households that made up the BHPS. See **www.understandingsociety.ac.uk** (accessed 9 January 2015).	Household organization; labour market behaviour; income and wealth; housing; health; socio-economic values.
Millennium Cohort Study	Study of 19,000 babies and their families born between 1 September 2000 and 31 August 2001 in England and Wales, and between 22 November 2000 and 11 January 2002 in Scotland and Northern Ireland. Data were collected by interview with parents when babies were 9 months and around 3 years old. Since then, surveys have been conducted at ages 5 and 7 years old. See **www.cls.ioe.ac.uk** (accessed 9 January 2015).	Continuity and change in each child's family and its parenting environment; important aspects of the child's development.

The UK Data Archive

- stores quantitative data from previous studies
- housed at the University of Essex
- online catalogue available at:
  - *http://**www.dataservice.ac.uk***
- documentation for each study
  - *topic, method, sample, sponsors, publications*
- download and order datasets

The Joint Economic and Social Data Archive

- stores quantitative data from surveys and statistical trends
- housed at the Higher School of Economics
- online catalogue available at:
  - http://sophist.hse.ru/eng/
- documentation for each study
  - *topic, method, sample, sponsors, publications*
- datasets available for free

Official statistics

- Collected by agencies of the state, in the course of their business
  - *e.g. the Employment Service compiles data for the level of unemployment*
- Advantages over quantitative data from surveys
  - *reduced time and cost*
  - *no problem of reactivity*
  - *cross-sectional and longitudinal analysis*
  - *cross-cultural analysis*

Disadvantages of official statistics

- Only reveal ‘tip of the iceberg’
  - *the ‘dark figure’ of unrecorded events*
  - *unemployed people who do not claim benefits are not officially listed as unemployed*

- The process used for data collection needs interpretation
  - *dubious measurement validity*

Problems with the *reliability* and *validity* of official statistics

- Reliability
  - *definitions, categories and allocated resources change over time*
  - *reflects priorities of agencies/organizations*
  - *e.g. changing definitions of crime*

- Validity
  - *variation may be caused by factors not studied by official reports*
  - *the ecological fallacy*

What is ‘the ecological fallacy’?

It is the error of assuming that inferences about individuals can be made from findings relating to aggregate data.

For example, official statistics might demonstrate a higher incidence of crime in regions with high concentrations of ethnic minorities but the members of the minority groups might not be responsible for the high level of crime.

Condemning official statistics

The widespread criticism of official statistics and their uses has led to their being largely ignored by social researchers.

In any event, they are not tailored to the needs of social researchers.

Resurrecting official statistics

Some official statistics – like population census data – are accurate by any set of criteria

To reject them because they contain errors is silly, since all measurement in social research is error-prone

The data is gathered ‘unobtrusively’, which means it is free from ‘reactive’ effects.

What are unobtrusive methods?

Webb et al. (1966) distinguish four main types:

- Physical Traces
- Archive materials
- Simple observation
- Contrived observation

Big data

Usually taken to refer to extremely large sources of data that are not immediately amenable to conventional ways of handling them.

It is often focussed on social media in social research, but is used to look at consumer behaviour by retailers.

Concerns that full potential of big data is not utilised.

The sources are non-reactive.