DMP title

Project Name Identity, Media and Consumer Behavior (Internal Funds DMP) - DMP title
Project Identifier C14/21/012
Principal Investigator / Researcher Mattia Nardotto
Institution KU Leuven

1. General Information

Name of the project lead (PI)
Mattia Nardotto

Internal Funds Project number & title
C14/21/012

2. Data description

2.1. Will you generate/collect new data and/or make use of existing data?

- Reuse existing data

2.2. What data will you collect, generate or reuse? Describe the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a numbered list or table and per objective of the project.

Scanner data of household purchases	.dta	Approx 1.8TB	Collected by grocery chain and transfered to the researcher on enchrypted hard disk
Twitter public tweets	.csv	100GB	Collected by Twitter and transfered to the researcher
Newspaper data	.csv	500MB-1GB	Downloaded from the National Archives (publicly available data)
Census data	.csv	less than 100MB	Downloaded from the UK's Office of National Statistics (publicly available data)
Electoral data (general elections of 2015 and Brexit referendum)	.csv	less than 100MB	Downloaed from the National archives (publicly available data)

3. Ethical and legal issues

3.1. Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to the file in KU Leuven's Record of Processing Activities. Be aware that registering the fact that you process personal data is a legal obligation.

The project makes use of:

- Data on purchases of households at the supermarket. Data has been anonymized before being transferred to the PI who has no information regarding the identity of the households.

- Data on public tweets related to the Brexit referendum published by users of Twitter.

The use of these datasets have been approved by the PRET and by the compliance monitoring team (ZK: ZKE0423)

The other datasets are publicly available data.

3.2. Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s).

No dataset that poses ethical issue is used.

3.3. Does your research possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

No

3.4. Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions regarding

reuse and sharing are in place?

No

4. Documentation and metadata

4.1. What documentation will be provided to enable understanding and reuse of the data collected/generated in this project?

The PI provides a detailed description of the variables used in the research in the form of a dictionary and all code for the statistical analysis is going to be published as well.

4.2. Will a metadata standard be used? If so, describe in detail which standard will be used. If not, state in detail which metadata will be created to make the data easy/easier to find and reuse.

No standard is going to be used.

Note: Being proprietary, the scanner data on purchases cannot be re-used (unless the company agrees to access it to another research team).

For the all dataset, I will create a custom metadata file using MS Excel and I will record:

Name of the resource	E.g., UK Census 2011
The type of content of the resource	e.g., Demographic characteristics of census areas
The entity that produced the resource	e.g., UK Office of National Statistics
Date of creation	
Format	e.g., .csv files
Language	e.g., English
Link to the resourse (if applicable)	e.g., https://www.ons.gov.uk/census/2011census

for some of the resources the list of entries for the table can be extended if needed.

5. Data storage and backup during the project

5.1. Where will the data be stored?

The publicly available datasets (Newspapers' headlines, Census, electoral data) have been downloaed and stored in the PC of the PI and a backup copy is going to be placed on the KUL OneDrive.

The data on households' purchases and the Twitter data have been stored in a encrypted hard drive which requires the manual input of a 16-digit key to be accessed. Only aggregated data obtained from the original data is then copied to the PC of the PI and on the KUL OneDrive. To give an example of aggregate data, the purchases of individual households are aggregated at the store/day level so that it is no longer possible to identify or track a household consumption profile. Similarly, Twitter data is aggregated by content at the level of the county or of the region in each day and thus also in this case it is no longer possible to identify or track twitter patterns of individual users.

5.2. How will the data be backed up?

See previous answer:

- The data on households' purchases and the Twitter data have been stored in a encrypted hard drive which requires the manual input of a 16-digit key to be accessed. Only aggregated data obtained from the original data is then copied to the PC of the PI and on the KUL OneDrive.

- The publicly available datasets (Newspapers' headlines, Census, electoral data) have been downloaed and stored in the PC of the PI and a backup copy is going to be placed on the KUL OneDrive.

5.3. Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

Yes, no problem of storage.

5.4. What are the expected costs for data storage and backup during the project? How will these costs be covered?

The cost for data storage is low. The encrypted hard drive has been bought (2TB hard drive is

sufficient) and no expensive server is needed so everything is going to fall within the running ZAP budget.

5.5. Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

See answer above: The data on households' purchases and the Twitter data have been stored in a encrypted hard drive which requires the manual input of a 16-digit key to be accessed. Only aggregated data obtained from the original data is then copied to the PC of the PI and on the KUL OneDrive.

The data policy and protection has been already approved by PRET and by the compliance monitoring team (ZK: ZKE0423)

6. Data preservation after the end of the project

6.1. Which data will be retained for the expected 10 year period after the end of the project? If only a selection of the data can/will be preserved, clearly state why this is the case (legal or contractual restrictions, physical preservation issues, ...).

The research team will preserve all data used in the project for the expected 10 years.

6.2. Where will these data be archived (= stored for the long term)?

The data will be stored by the research team for 10 years after the end of the project. We will make use of the encrypted hard drives that we already have, and we are going to buy a secondary hard drive to make sure that a copy is also available in case one of the two is corrupted over the years.

6.3. What are the expected costs for data preservation during these 10 years? How will the costs be covered?

The external hard drive we are currently using is **iStorage diskAshur Pro2 2TB.** It's cost is fairly reasonable and sustainable for the PI (i.e., around 300 euro).

7. Data sharing and re-use

7.1. Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions or because of IP potential)?

The grocery data is proprietary and can be shared only with the agreement of the owner.

The Twitter data can be obtained by other researchers either by scraping the platform using the API or by buying the same data from the company. Also in this case, the data cannot be shared by the PI without the consent of Twitter.

All other datasets, being public, can be obtained by anybody without restrictions.

7.2. Which data will be made available after the end of the project?

Only the public data can be made available to everybody.

7.3. Where/how will the data be made available for reuse?

- In an Open Access repository

(of course it applies only to the data that can be made publicly available)

7.4. When will the data be made available?

- Upon publication of the research results

7.5. Who will be able to access the data and under what conditions?

The PI and the research team (e.g., the PhD student involved in the project).

7.6. What are the expected costs for data sharing? How will these costs be covered?

No cost.

8. Responsibilities

8.1. Who will be responsible for the data documentation & metadata?

The PI

8.2. Who will be responsible for data storage & back up during the project?
The PI

8.3. Who will be responsible for ensuring data preservation and sharing?
The PI

8.4. Who bears the end responsibility for updating & implementing this DMP?
The end responsibility for updating and implementing the DMP is with the supervisor (promotor).