Data Management Plan – *"Network Industries: Measuring how Network Externalities Matter for Pricing and Entry"*

FWO PhD Scholarship for Fundamental Research (11D1522N)

ADMIN DETAILS

Project Name:	**“**Network Industries: Measuring how Network Externalities Matter for Pricing and Entry"
Grant Title:	11D1522N
Principal Investigator / Researcher:	William Harry Burton
Institution:	KU Leuven

1. GENERAL INFORMATION

Name applicant

William Harry BURTON

FWO Project Number & Title

"Network Industries: Measuring How Network Externalities Matter for Pricing and Entry" (11D1522N)

Affiliation

- KU Leuven

2. DATA DESCRIPTION

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

- Generate new data
- Reuse existing data

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

Please see the table below. I will refer to this as Table 1 of Section 2 throughout the data management plan.

Data set name	Source of the data set	Work packages using this data	Origin of the data points	Content	Timespan of the data	Type of data	File format(s)	Size	Access	Shareable	Data set status
Demand Data	Licensed from IATA ("Market Intelligence Service" data).	- WP 1 - Possibly WP 3	IATA retrieves this data from a global database of sold aviation tickets. This data is already aggregate information, and therefore not subject to any GDPR regulations.	Price and quantity data regarding airplane tickets starting, ending or transferin g in Sweden.	- Jan. 2017 - Jun. 2017 - Jan. 2018 - Jun. 2018	Numerical and string variables.	.csv .RData	200 MB	Proprietar y data	No, due to licensing of the original data.	Complete
Flight Movement	Downloaded from Eurocontrol R&D data archive.	- WP 1 - WP 2 - WP 3	Filed and actual flight plans of aircrafts operating in Europe. These flight plans are collected by Eurocontrol.	Departure and arrival times and locations, aircraft registratio n, operating company, flight distance	2015 - 2020; for each year except 2020: flight plans from Mar., Jun., Sep. and Dec. For 2020: only Mar.	Numerical and string variables.	.csv .RData	1 GB	Open-access (after registratio n)	Already open-access via Eurocontrol R&D data archive	Complete (gradual addition of new months by Eurocont rol)

Airline Metadata	Manual collection from Wikipedia and various other online sources	- WP 1 - WP 2 - Possibly WP 3	Public company information, as well as public administrative data	Name, ownership structure, country of origin, IATA and ICAO airline codes, number of aircrafts owned	Ground truth in January 2017, and any subsequent changes.	Numerical and string variables.	.csv	100 MB	Open-access	Will be made publicly available once KU Leuven has approved its open-access release (possible research valorisation).	Complete (gradual updating)
Airport Metadata	Manual collection from Wikipedia, various airport websites, and the Data Set "Flight Movement"	- WP 1 - WP 2 - Possibly WP 3	Public company information, as well as public administrative data	Name, location, infrastructure information (e.g. number of runways, number of terminals/ gates, ...)	Ground truth in January 2017, and any subsequent changes.	Numerical and string variables.	.csv	100 MB	Open-access	Will be made publicly available once KU Leuven has approved its open-access release (possible research valorisation).	Complete (gradual updating)
Subsidy Information	Manual collection from Wikipedia, news reporting, filings with the European Commission and	- WP 2 - Possibly WP 3	Public administrative data regarding the subsidisation of airlines and/or airports.	Time, Subsidy-granting institution, subsidy-receiving	So far data from 01/2018 till 12/2021 has been	Numerical and string variables.	.csv	500 MB	Open-access / Public Records	Will be made publicly available once KU Leuven has	In collection

	national competition authorities.			institution, subsidy amount, subsidy conditions	collected. Expansion until 01/2000 is intended.					approved its open-access release (possible research valorisation).	
Simulated Demand Data	Simulation of a demand data set according to the methodology outlined in the proposal.	- WP 3	Simulation of a structural model of demand and competition in the airline industry using code written in the Python programming language.	Simulated price and quantity demand information for hypothetical products offered by hypothetical airlines.	Several hypothetical time periods (discrete time simulation)	Numerical and string variables	.csv	100 MB	Currently code is in private development mode.	Code will be made publicly available on Github. This allows easy replication of the data set. Simulation metadata will also be provided.	In development.

3. LEGAL AND ETHICAL ISSUES

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for Research and Public Service Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

- No

I will not be using any data that can identify an individual. One may think that the data set “Demand Data” listed in Table 1 in Section 2 is personal data. However, the data is aggregate market-level data with each data point essentially representing a total number of sold tickets in one month in a market. It is therefore not person-specific. Furthermore, the data provider (i.e. IATA) has taken extensive steps to mask (or remove) critical data points before the data is released to any licensee for analysis purposes.
On a final note, IATA (and their representatives) mentioned at no point during the licensing process that this was GDPR-relevant data, suggesting that they agree with the aggregate data assessment presented above.

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

Does your work 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?

- Yes

I am cautiously suggesting that there might be potential commercial exploitation of the data sets “Airline Metadata”, “Airport Metadata”, and “Subsidy Information” listed in Table 1 in Section 2. The first two data sets are very similar to ones offered by for-profit firms. The data itself is not proprietary, but these firms charge customers for the manual labour that is involved in collecting it.
I will be contacting the KU Leuven Intellectual Property Department soon to seek guidance on the “releasability” of these data sets, and whether KU Leuven as the IP holder allows the above mentioned data sets to be made publicly available under the "Creative Commons Attribution Share-Alike” (CC-BY-SA) license.

Any data set or software package, which is licensed or proprietary, will of course not be published.

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 are in place?

- Yes

The data set "Demand Data" listed in Table 1 in Section 2 is licensed from IATA. The licensing agreement forbids the licensees (i.e. co-authors and myself) from publishing, sharing or distributing the data in any shape or form. Research findings may be made public, conditional on not revealing individual data points. The licensing agreement however allows for the data to remain on a secure and access-restricted storage server. Individuals wishing to reproduce the results that stem from this data set must therefore license the data from IATA themselves in order to be successful in the reproduction of results.

4. DOCUMENTATION AND METADATA

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

1. Conditional on a particular work package, each data set that is used by the respective work package will receive its own data dictionary. This dictionary will describe various aspects, amongst other things: the definition of each variable as well as its underlying meaning, its original source, unit of measurement (if applicable), whether it is a "raw" variable (i.e. if this variable is contained in the original and unprocessed data set) or a "processed" variable (i.e. a raw variable that is somehow transformed by code), etc. Furthermore, each data dictionary will feature sections explaining which scripts of code needs to be run in order to re-create processed variables, as well as sections explaining which scripts of code use which variables. All these dictionaries will be collected in a folder entitled “data_dictionaries”. All this will lower the burden of reproducibility enormously.

2. All sets of simulated data will receive a unique identifier. It will be present as a variable in each row of any simulated data set (each row representing one simulated observation). This identifier will then be listed in a separate file entitled “data_simulation_setups.md”, which will describe the various parameters originally used to generate the data set. It will also contain the specific seeds that were set while the data was generated/simulated, in order to allow those interested to reproduce the “randomness” in the simulated data set.

3. The methodology applied to any manually collected data set will be documented in a separate file entitled “data_collection_methodology.md”. This will those interested to understand how the data set was compiled. Furthermore, the source of each data point will either be preserved (e.g. for text files that are easily storable, such as PDFs) or referenced correctly (e.g. websites and their date of retrieval). Furthermore, manually collected data sets will feature a flag variable (0/1 encoding) as well as comments on an observation-level, allowing the individual who recorded/collected the data point to express any concerns that arose during the manual collection process. For example, if conflicting information about an observation was found from two different credible sources, this data point will be flagged and an appropriate comment will be written. This will allow those interested to understand why this data point was included/excluded.

4. For each work package, any sample restrictions that were applied to data sets (e.g. outlier removal) will be documented in a separate file entitled “data_sample_restrictions.md”. This file will also contain documentation on the flow of sample-restricted data sets: which scripts of code use which sample-restricted data sets, and which results therefore depend on which sample-restricted data sets.

5. All lines of code will be commented, describing exactly what each step does. At the top of each script, there will also be a header listing the file name, code author, date of last revision, as well as brief description of what the code contained in the script does.

6. All scripts of code will be version-controlled using the version-control software Git, allowing those interested to see and understand the entire development process of the code.

7. Each work package will contain a separate file entitled “project_necessary_steps.md”, specifying the exact steps (i.e. the sequence of scripts) which are necessary to reproduce any intermediate and/or final results (e.g. summary statistic tables, regression results, etc.). These documented steps will start with the raw data sets, explaining which processing scripts to run, before turning to the sequence of the analysis scripts.

8. Each work package will contain the same folder structure. On the top level there will be (at least) three folders entitled "data", "code", and "documentation". The data folder will contain "raw", "processed", and "final" folders, allowing for the appropriate grouping of data sets. The code folder will contain "build", "analysis" and "graph" folders, again allowing for the appropriate grouping of scripts. The documentation folder will feature no separate folders, but document names will be chosen according to their content.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- Yes

I will be using the metadata standard “DDI-Codebook 2.5”, which is the at the time of writing this initial DMP the most recent version. In my opinion, this metadata standard gives me the greatest flexibility in terms of available fields/elements. This will allow me to build a machine-readable document that will complement my already extensive data set documentation, as I have outlined in other sections of this DMP.

5. DATA STORAGE AND BACKUP DURING THE FWO PROJECT

Where will the data be stored?

1. Any raw data will be stored in the following four locations: locally on my work computer (in a password-protected folder), on my encrypted external SSD, on Dropbox for collaboration with co-authors, and in my personal network drive (OneDrive provided through KU Leuven). Please note that both Dropbox and my KU Leuven OneDrive are also access-restricted, thus disallowing unauthorised access.

2. Any code that is developed will be stored in the following three locations: locally on my work computer (in a password-protected folder), on Dropbox for collaboration with co-authors, and on a remote repository on Github (which is set to private until the code is ready to be shared).

How is backup of the data provided?

1. Any raw data sets will be backed up to the following two locations on a weekly basis: Dropbox, as well as my personal network drive (OneDrive provided through KU Leuven). These are managed services, which also allow access to previous/"historical" versions of the data sets.

2. By using the version-control software Git, any (committed) changes to the code base of each work project will automatically backed up to a local .git file. Furthermore, I make use of remote (private) repositories in the following three locations to which changes will be pushed at the end of each development day: Github, Dropbox, and my encrypted external SSD.

3. Processed data sets will not be backed up in a systematic way as reproducing them is very easy due to the detailed data and code documentation, as well as the backed-up raw data and scripts.

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

1. My raw data sets do not require too much space as outlined in . Currently, my Dropbox features 2TB of space of which only 0.2% are used. Furthermore, my KU Leuven OneDrive also features 2TB of space of which currently only 1GB is used.

2. My code scripts are not very demanding in terms of storage space, and therefore size constraints can be disregarded. Furthermore, Github does not restrict the size of neither public nor private code repositories.

What are the expected costs for data storage and back up during the project? How will these costs be covered?

1. Github is free of charge.

2. OneDrive is provided through KU Leuven, and free of charge.

3. The Dropbox Plus fee is currently just below EUR 120 per year (EUR 9.99 per month), and can therefore be easily covered by the annual FWO benchfee.

4. My external SSD is also used for private purposes, and therefore any costs in case of repair or replacement are covered by me personally.

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

1. All Github repositories are set to private, and are therefore access-restricted. This “private” setting will be changed to “public” once the code is ready to be shared. However, even then modifications will only be possible through so-called “pull-requests”, which in turn need to be approved by the repository owner. This protects the code base of any project against any malicious attempts to change the code.

2. With the exception of one data set, all of the data sets contain (easily accessible) public information. This allows one to focus on preventing modifications by unauthorised people. This is done by only granting write-access to data-collection collaborators, with the additional restriction that final data set changes need to be reviewed ex-ante by a trusted individual. However with frequent backups, even the damage of malicious changes by entrusted individuals is also limited.

3. The data set which requires the most secure storage is the “Demand Data” data set listed in Table 1 in Section 2. This data set is currently stored within the PIN-protected file vault provided by Dropbox. Therefore, three levels of security (restricted access, strong user passwords, and a PIN) are implemented for this data set.

4. All users of the data sets have agreed to a strong password policy for their accounts.

6. DATA PRESERVATION AFTER THE FWO PROJECT

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

All data sets listed in Table 1 in Section 2 and their documentation will be retained for a minimum of 10 years. Data sets which are however derivatives of these raw data sets (e.g. transformed through code) will not be stored as they can be easily reproduced by following the detailed documentation for reproduction. Furthermore, it would create unnecessary data storage costs.

All code scripts will also be retained for a minimum of 10 years.

One exception might apply to the “Demand Data” data set from Table 1 in Section 2. As the data is licensed from IATA, the licensing agreement includes a clause which stipulates if a dispute were to occur and this dispute cannot be resolved, that one must delete all versions of the data. However, the data itself is not lost as IATA is a well-established organisation, which in my expectation will not be dissolved in the next ten years. Those who are interested in (re-)licensing the data, can then do so from IATA.

Where will the data be archived (= stored for the longer term)?

All raw data sets, their documentation, as well as code scripts will be archived on KU Leuven's central storage servers for at least ten years. Access to the “Demand Data” from IATA will however be restricted in accordance with the licensing agreement. The other data sets mentioned in Table 1 in Section 2 as well as their documentation will be freely accessible (conditional on KU Leuven approval). If approval is granted, those data sets will also be shared via / uploaded to Zonedo in a .csv format, as well as the appropriate documentation. A further data, documentation, and code archive will be provided on my personal website (williamburton.eu).

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

All data sets, documentation, and code scripts will be stored on KU Leuven servers. I do not expect this to cost a significant amount (internal KU Leuven documents suggest 50 EUR per year for 100 GB). Thus, the FWO bench fee will be sufficient to cover this ten-year fee.

Github and Zenodo are currently free of charge, and thus provide archiving solutions at no cost.

My personal website (williamburton.eu) is paid for by myself.

As the data sets, the documentation, as well as code scripts are kept in good condition at all stages of the project, preparing the data for the aforementioned archives will only result in a few hours of labour.

7. DATA SHARING AND REUSE

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)?

- Yes.

The data set "Demand Data" listed in Table 1 in Section 2 is licensed from IATA. The licensing agreement forbids the licensees (i.e. co-authors and myself) from publishing, sharing or distributing the data in any shape or form. Research findings may be made public, conditional on not revealing individual data points. The licensing agreement however allows for the data to remain on a secure and access-restricted storage server. Individuals wishing to reproduce the results that stem from this data set must therefore license the data from IATA themselves in order to be successful in the reproduction of results.

The sharing of the other data sets listed in Table 1 in Section 2 may be prohibited by KU Leuven's Intellectual Property Department due to the potential commercial exploitation mentioned in Section 3.

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

All manually collected data sets will be made available in a .csv format under a CC-BY-SA license, conditional on receiving approval from KU Leuven's Intellectual Property Department. How access is provided is discussed below.

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

- In an Open Access repository
- Upon request by mail
- Other

1. All manually created data sets will be made available in a .csv format on Zenodo, accompanied by the appropriate documentation and metadata. Furthermore, they will be uploaded to my personal website (i.e. williamburton.eu), and be kept there as long as my academic career continues, but at least a minimum of ten years after the end of my projects.

2. All scripts will be made available via Github by making use of public repositories.

When will the data be made available?

- Immediately after the end of the project

The data, which I am allowed to published, will be made available on a rolling basis. As soon as a sufficient number of entries has been added, or a substantial amount of previous entries has been

updated, I will publish a new version of the data set. This can also occur prior to the end of a project. “Pre-releasing” the data is in my opinion also good for the progress of my projects as others can contribute data points or point towards mistakes I have made.

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

All data sets I am allowed to share will be provided under a CC-BY-SA license. All code will be provided under the MIT license.

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

Github and Zenodo are currently free of charge. I do not expect this to change in the (nearby) future. Should this change within the funding time frame, some money of the FWO bench fee will be used (if this purpose is covered), or I will use personal funds as I do not expect significant costs for data sharing.

8. RESPONSIBILITIES

Who will be responsible for data documentation & metadata?

William Harry BURTON

Who will be responsible for data storage & back up during the project?

William Harry BURTON

Who will be responsible for ensuring data preservation and reuse ?

William Harry BURTON

Who bears the end responsibility for updating & implementing this DMP?

The PI bears the end responsibility of updating & implementing this DMP.