DMP title

Project Name Opening access to economic data to prevent tobacco related diseases in Africa

Project Identifier PROP0715E

Grant Title 108098-001

Principal Investigator / **Researcher** Corne Van Walbeek

Project Data Contact Lynn Woolfrey, +27216505707, lynn.woolfrey@uct.ac.za

Description The purpose of this project is to demonstrate that data for tobacco control research in sub-Saharan Africa can be collected and distributed from an Open Data platform and used in policy relevant research activities. The platform and data will improve the capacity for tobacco control research in key sub-Saharan African countries, and help develop a continent-wide research approach to tobacco control.

Data Collection

What data will you collect or create?

https://drive.google.com/open?id=0B8r2zPDiNDLfazVTUXhGR3VEdXM

How will the data be collected or created?

Data Collection methods:

1. Desk-based search of official websites of project countries:

This will involve searching websites of government departments of selected Sub-Saharan African countries for administrative data collected by these departments which relate to tobacco production or usage.

For example. These websites will be scraped for data on:

a. Tobacco production data (Departments of Agriculture)

b. Prevalence of tobacco-related diseases, and tobacco-related morbidity and mortality (Departments of Health)

c. Tobacco taxation (Internal Revenue Services)

d. Tobacco products manufacturing, tobacco imports/exports (Departments of Trade and Industry)

National Statistics Agencies (NSAs) websites are another useful place to find administrative data. In South Africa unit record administrative data from departments, repackaged as research datasets, are shared by the NSA

If data collection instruments (administrative forms) used to collect the data are available on these sites they can provide useful information on data fields and subject categories for final datasets

We will follow this up with a study of useful variables on tobacco from country surveys.

Again, an initial desk-based study will allow us to download public use data from project countries, and interrogate these for tobacco-related variables. From this a "question bank" will be created of useful variables and the datasets where these can be found.

2. Desk-based search of industry websites

The second component of our desk-based research will involve examing online records of the tobacco industry. From these we hope to obtain data on:

Cost of tobacco production, and profits, in the industry, prices of raw tobacco and tobacco products, salaries, capital and foreign investment, mergers and acquisitions, advertising spend, and regulations in the industry

3. Approaches to data holders

Our desk search may reveal the existence of datasets with a tobacco data component but which are not in the public domain. In these cases we will approach NSAs or the relevant research projects in the project countries to release this data and allow the Project to host this on their Open Data portal. This may be a fraught process but any challenges and successes can be written up to inform our future work.

4. Own surveys

The project has already crowd-sourced data on current prices of tobacco products in two project countries. This may be expanded during the course of the project to all project countries.

Documentation and Metadata

What documentation and metadata will accompany the data?

Supporting documents

These documents will be shared with the data files, where available. Forms used for collecting administrative data will be shared with administrative datasets. Data collection instruments (questionnaires, diaries) will be made available with the survey data. Code lists used in collecting the data will also be provided. Final reports from data collection projects will form part of datasets, where available.

Metadata

Each dataset will have a metadata record to help data users analyse the data. This metadata record will be created during examination of the data and data collection instruments. It will include information gathered on the dataset during the data collection process. The latter is often useful for those analysing the data. Issues around data quality will form part of the metadata record. Metadata will be created according to the Data Documentation Initiative (DDI) international metadata standard, using Nesstar Publisher, which is free data markup software for the creation of XML compliant metadata according to the DDI standard.

Ethics and Legal Compliance

How will you manage any ethical issues?

The administrative data we will collect will mostly be in the public domain, in the form of reports and other records from government departments. The survey data we will collect will be anonymised data already shared with researchers, although not always online. The industry data will be data made available to shareholders and the public. We are adding value by bringing these sources together and providing a means for researchers to easily discover and download these data.

However, we will endeavour to make data available that is not yet in the public domain. In these cases, we will ensure that:

a. We have the necessary permissions from data owners make these data open.

b. The data is suitably anonymised, to protect respondent confidentiality and privacy

c. We take national laws on sharing data across borders into account. Where such restrictions exist, we will be unable to host this data.

d. We work with all stakeholders to ensure agreement on what will be shared, how, and with whom.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

The government data we will collect is not subject to IPR. The tobacco industry data we will collect will be data made public by the industry, which does not publish information that would compromise their IP rights. However, we will check each traunch of data we obtain, to ensure we have permission to pass the data on to third parties.

Storage and Backup

How will the data be stored and backed up during the research?

The data would be stored on a server managed and backed up by the University of Cape Town's Commerce IT Department. Curation of the data will be the responsibility of DataFirst's Research Data Service. DataFirst is a technical partner on the Project. Each preservation dataset will consist of data files, document files, metadata files, and any programme files used in creating the data files. Data Service staff will be responsible for adding data updates to datasets. We will also handle version control to ensure the most recent and accurate data files are published, and earlier versions are available for verification or replication of research which may cite earlier versions of the data.

How will you manage access and security?

Access to the server hosting the preservation datasets will be password controlled. Passwords will be allocated by the Commerce IT manager only to Data Service staff. Server software will monitor data security and integrity.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Criteria for preservation will be:

a. Data is tobacco-related

b. Data covers project countries

c. Data is accurate and reliable (we will undertake quality audits to determine this)

d. Data is unit record data not available in another dataset in the collection

e. Data is not readily available from another repository

Retention:

It is difficult to predict what data has long-term value. Our policy will be to store unit record tobacco data indefinitely. As these datasets grow so will their value over time. Time series data continue to be useful for economic and health policy research in the long term.

Sharing:

Because we aim to establish an Open Data portal, all data retained/preserved will also be shared. The Project's policy is aligned to DataFirst's policy: We do not archive data which cannot be shared with researchers in some form and at some access level.

What is the long-term preservation plan for the dataset?

There will be numerous datasets. Our long term preservation plan for the Project's data holdings depends on the sustainability of DataFirst's Research Data Service. The service was established in 2001 and is a unit at the University of Cape Town, a well-funded and well-established university in South Africa. Our sustainability prospects are therefore good.

Data Sharing

How will you share the data?

The data will be shared as discrete datasets (by country, year, data source). DataFirst hosts and shares data via an online dissemination platform, based on the National Data Archive Open Source software developed by the World Bank's Development Data Group

The platform provides a number of data access options. The Project's data will be shared as Public Use data. That is, researchers will need to register on our site and say for what purpose they will use the data, but access will be immediate and automatic, with no vetting of use. The usage information we collect will be to support service improvements.

The data will be shared in a number of formats, includind excel spreadsheet, and data files in the commonly used statistical analysis programmes (SPSS, Stata). We will also make the data available as .csv files, in line with Open Data requirements.

Are any restrictions on data sharing required?

We aim to share the tobacco data we collect as public access data. We do not aim to support research-use only requirements, as this is counter to Open Data principles. Policy research, academic research, business analysis, and private sector innovation all need good data, and countries benefit from informed decision-making in all these spheres.

Responsibilities and Resources

Who will be responsible for data management?

The Manager of DataFirst's Research Data Service will be responsible for curating the Project's data. This is in line with the project proposal for DataFirst to be funded to provide technical support. DataFirst's Manager has 25 years' experience in managing research data and working with data users. Experience from undertaking data rescue projects in South Africa will also be useful in assisting with data collection activities.

What resources will you require to deliver your plan?

Funding for data collection as been budgeted for in the Project. This may need to include funding to travel to project countries and negotiate with data collectors in government and academia to release their data, and allow its reuse. Funding has been provided for a Project Manager. The Project Manager is responsible for conducting online data audits and downloading data, and populating the database which DataFirst's Manager will curate. This will be a time- and labour- intensive task and more staff hours may need funded for this.