Research Ideas and Outcomes 3: e14837
doi: 10.3897/rio.3.e14837

Data Management Plan

Data Management Plan: Opening access to economic data to prevent tobacco related diseases in Africa

Lynn Woolfrey ‡

‡ University of Cape Town, Cape Town, South Africa

Corresponding author: Lynn Woolfrey (lynn.woolfrey@uct.ac.za)

Reviewable v1

Received: 03 Jul 2017 | Published: 24 Jul 2017

Citation: Woolfrey L (2017) Data Management Plan: Opening access to economic data to prevent tobacco related diseases in Africa. Research Ideas and Outcomes 3: e14837. https://doi.org/10.3897/rio.3.e14837

Abstract

The purpose of this project is to demonstrate that tobacco-related data from selected Africa countries can be collected and distributed from an Open Data platform. 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.

Keywords

DMP, data management, tobacco, health, tobacco related disease, economics

Data Collection

What data will you collect or create?

See Fig. 1.

Tobacco Data in Africa Project (IDRC Project 108131) Data Inventory 2016								
Data Source	**Data Description**	**Data Type**	**Data Collection Instruments**	**File Format**	**File Size**	**Data Discovery Issues**	**Data Access Issues**	**Data Quality Issues**
Dept of Agriculture	**Cultivation area of tobacco crops Total tobacco production Farm gate prices of tobacco**	**Administrative records**	**Administrative forms**	**MSWord, pdf, excel spreadsheets**	**Small (Kilobytes)**	**Very little data is listed online. Absence of discovery metadata.**	**Negotiating permissions. No tradition of sharing government administrative data. Data is locked away in inaccessible formats, eg e.g. non-digital, pdf.**	**Accuracy - issues around data collection. Reliability - trustworthy sources? Sources contradict each other Comparability - over time periods and between domains Timeliness - time lags in collection/collation of data Interpretability - almost no data documentation, metadata Big Data issues - storage and throughput - Administrative database could grow to many petabytes. Need computing power for data linking e.g. admin and panel data**
Dept of Health	Prevalence of tobacco-related diseases; Disease burden (annual deaths); Exposure to second-hand smoking (% children exposed); Tobacco use (cigarettes/other tobacco products/all); Legal environment - regulations on health warnings, smoke-free areas	Administrative records and survey data	Administrative forms	Access databases, excel spreadsheets	**Small (Kilobytes)**	Very little data is listed online. Absence of discovery metadata.	Negotiating permissions. No tradition of sharing government administrative data. Data is locked away in inaccessible formats, eg e.g. non-digital, pdf.	Accuracy - issues around data collection. Reliability - trustworthy sources? Sources contradict each other Comparability - over time periods and between domains Timeliness - time lags in collection/collation of data Interpretability - almost no data documentation, metadata Big Data issues - storage and throughput - Administrative database could grow to many petabytes. Need computing power for data linking e.g. admin and panel data
Dept of Trade and Industry	**Tobacco products manufacturing data; Tobacco leaves - import/export data (volume); Tobacco products - import/export data (volume); Major partners in tobacco trade**	**Administrative records**	**Administrative forms**	**MSWord, pdf, excel spreadsheets**	**Small (Kilobytes)**	**Very little data is listed online. Absence of discovery metadata.**	**Negotiating permissions. No tradition of sharing government administrative data. Data is locked away in inaccessible formats, eg e.g. non-digital, pdf.**	**Data quality issues: Accuracy - issues around data collection. Reliability - trustworthy sources? Sources contradict each other Comparability - over time periods and between domains Timeliness - time lags in collection/collation of data Interpretability - almost no data documentation, metadata Big Data issues - storage and throughput - Administrative database could grow to many petabytes. Need computing power for data linking e.g. admin and panel data**
Revenue Service	Tobacco tax data Tax structure/rate (historical) Tax base Tax revenue by type of tax Total tax revenue Average excise tax per cigarette pack (total tobacco excise revenue/total sales) Average total tax per cigarette pack (total tobacco tax revenue/total sales) Average excise tax per other unit of tobacco (total tobacco excise revenue/total sales) Average total tax per other unit of tobacco (total tobacco tax revenue/total sales)	Administrative records	Administrative forms	MSWord, pdf, excel spreadsheets	**Small (Kilobytes)**	Very little data is listed online. Absence of discovery metadata.	Negotiating permissions. No tradition of sharing government administrative data. Data is locked away in inaccessible formats, eg e.g. non-digital, pdf.	Data quality issues: Accuracy - issues around data collection. Reliability - trustworthy sources? Sources contradict each other Comparability - over time periods and between domains Timeliness - time lags in collection/collation of data Interpretability - almost no data documentation, metadata Big Data issues - storage and throughput - Administrative database could grow to many petabytes. Need computing power for data linking e.g. admin and panel data
Tobacco Industry	**Cost of tobacco production Prices of raw tobacco Price per cigarette pack (most popular brand/imported brands/all) Average price per standard cigarette pack (20 cigarettes) (total revenue and total sales) Prices of other tobacco products (by kind/brand name/unit) Salaries in tobacco industry Capital investment in tobacco industry Advertising spend of tobacco industry Profit of tobacco industry - domestic firms Profit of tobacco industry - importing firms Income tax paid by tobacco industry Wholesale, retail margin of tobacco industry Market share by tobacco company Market share by brand/ most sold brands Production of tobacco products Sales of tobacco products Mergers and acquisitions of tobacco companies Foreign investment and (including regulation)**	**Administrative records**	**Annual and financial reports**	**MSWord, pdf, excel spreadsheets**	**Medium (Megabytes)**	**Very little data is listed online. Absence of discovery metadata.**	**Data is locked away in inaccessible formats, eg e.g. non-digital, pdf.**	**Accuracy of the data Reliability - trustworthy sources? Sources contradict each other Comparability - over time periods and between domains Timeliness - time lags in collection/collation of data Interpretability - almost no data documentation, metadata**
National Statistics Agencies (NSAs)	Data related to tobacco consumption and expenditure; Tobacco use - average age of initiation; Tobacco use - consumption intensity (e.g. no. cigarettes per day); Consumer Price Index (CPI) for cigarettes	Census/survey data	Questionnaires	Excel spreadsheets, SAS/SPSS/Stata files	Medium to large (Megabytes, Gigabytes	Data is locked away in inaccessible formats, eg e.g. non-digital, pdf. May need to travel to countries to obtain data	Negotiating project (and ongoing) permissions to share. Most African census data is not in the public/research domain. Only some African governments share their sample survey data.	Data quality issues: Accuracy - issues around data collection. Reliability - trustworthy sources? Sources contradict each other Comparability - over time periods and between domains Timeliness - time lags in collection/collation of data Interpretability - almost no data documentation, metadata

Figure 1. doi

Tobacco Data in Africa Project Data Inventory 2016. Original data available as Suppl. material 1.

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 These will include data on:

1. Tobacco production (Departments of Agriculture)
2. Prevalence of tobacco-related diseases, and tobacco-related morbidity and mortality (Departments of Health)
3. Tobacco taxation (Internal Revenue Services)
4. Tobacco products manufacturing, tobacco imports/exports (Departments of Trade and Industry)
5. Tobacco usage, from Surveys by National Statistics Agencies (NSAs). In South Africa unit record administrative data from government departments, repackaged as research datasets, are also 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 the data.

2. Desk-based search of websites of International Development Organisations

This desk-based study will allow us to discover tobacco data on project countries that has been collected by international organisations. The will include the websites of the international DHS Program and UN bodies such as the World Health Organisation (WHO). 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 third 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

We will create metadata on our platform for surveys already shared by others online. Our desk search may also reveal the existence of datasets with a tobacco data components but which are not in the public domain. In these cases we will approach 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. We will upload and share the metadata and data from these surveys.

Documentation and Metadata

What documentation and metadata will accompany the data?

Supporting documents

Data collection and data analysis documents will be shared along 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 also be shared, 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. This documentation can be an invaluable source of provenance and usage information for those analysing the data. Notes on 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:

1. We have the necessary permissions from data owners to make these data open.
2. The data is suitably anonymised, to protect respondent confidentiality and privacy
3. We take national laws on sharing data across borders into account. Where such restrictions exist, we will be unable to host this data.
4. 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 public records. We will therefore not publish information that would compromise any 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 provide tombstone citations to earlier versions for verification or replication of research which may cite these supercede 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:

1. Data is tobacco-related
2. Data covers project countries
3. Data is accurate and reliable (we will undertake quality audits to determine this)
4. Data is unit record data (not aggregated but available at the level at which it was collected)
5. 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 will 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 spreadsheets, 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 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. Skills learned 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.

Grant title

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

Hosting institution

University of Cape Town

Supplementary material

Suppl. material 1: Data quality issues: Accuracy - issues around data collection. doi

Authors: Lynn Woolfrey
Data type: Administrative records
Filename: 20170718-tobacco-data-inventory-1.xlsx - Download file (11.55 kb)