# FINAL TECHNICAL REPORT

PROJECT TITLE: EXPLORING THE OPPORTUNITIES AND CHALLENGES OF IMPLEMENTING OPEN RESEARCH STRATEGIES WITHIN DEVELOPMENT INSTITUTIONS

PROJECT CODE: 108131-006

INSTITUTION NAME: CENTRE FOR ANALYSIS AND FORECASTING (CAF)

IMPLEMENTING AGENCY: REAL-TIME ANALYTICS (RTA)

LOCATION: Hanoi, Thai Nguyen, Vinh Phuc, Thai Binh, Ha Tinh, Da Nang, Quang Nam, Dak Lak, Binh Duong, Ba Ria- Vung Tau, Ho Chi Minh, Tien Giang

AUTHORS:
- CENTRE FOR ANALYSIS AND FORECASTING (CAF)
- REAL-TIME ANALYTICS (RTA)

DATE OF REPORT: March 20th, 2017

***Executive summary***

Project 107145 “Strengthening the Economic Committee of the National Assembly in Vietnam” is one of the 8 IDRC-supported projects chosen to participate in the Open Sharing Pilot Project (Project 108131-006). Under the Project 107145, Real-Time Analytics’s main responsibility is to conduct a baseline survey and 3 followed-up surveys of 773 SMEs in 12 provinces. In addition to the quantiative surveys of 773 SMEs, RTA conducted 13 in-depth interviews on innovation-based startups.

The data collected through out the Project 107145 should be made available for public use. Nevertheless, how to share it openly and properly is not straightforward. There are a number of issues we have address, such as:

- We do we host the data? How long can the hosting be avalable? Who answers questions raised by data users regarding the data?
- What can be shared and what cannot be shared?
- What are the documentations of the data sets?

Participating in the Pilot Project helps to find answers to these questions. Dr. Le Dang Trung, Chief Economist of RTA, travelled to Canada twice to participate in the discussions with the other pilot projects. A Data Management Plan has been developed. The lessons learnt from the Pilot Project are valueable in many aspects. Now, RTA not only has the willingness to share the data, it also knows how to share properly.

***The research problem***

Project 107145 “Strengthening the Economic Committee of the National Assembly in Vietnam” has collected data via SME surveys and in-depth interviews. How can the data be shared openly to the public?

***Progress towards milestones***

The project has set 4 major milestones as the following:

- 1.1. Test and refine implementation guidelines for development research funders' open research data policies;
- 1.2. Examine the specific ethical and implementation issues in data sharing in the context of development research including issues of knowledge developed in development and indigenous contexts;
- 1.3. Initiate the development of a community of practice around open data management planning amongst Centre grantees; and

- 1.4. Build the capacity of a Centre grantee to manage and share open research data.

To fulfill the milestones, activities had been conducted along with the implementation of the Project 107145. There are two major lines of activities including: i) workshops in Ottawa and ii) implementation of the Data Management Plan. Details of the activities are provided below:

- Dr. Le Dang Trung completed the first trip to IDRC in Canada to learn about developing a Data Management Plan (DMP) in March 2016.

- After the trip, he and the implementing team had worked on the DMP and integrated it into the implementation of the Project 107145. During the development of the DMP, discussions with the Project investigator Professor Cameron Neylon were held.

- Dr. Le Dang Trung participated in the wrapping-up meeting in Ottawa in December 2016. The Data Management Plan was discussed and improved.

- After the second trip to Ottawa, the implementing team focused on preparing survey documentations including technical report and data set labels, ... to facilitate the sharing of data accordingly to the Data Management Plan

***Synthesis of research results and development outcomes***

Key results have been obtained as follows:

- The Data Management Plan for sharing data collected by the Project 107145 "Strengthening the Economic Committee of the National Assembly in Vietnam" has been developed

- Preparation for sharing the data has been completed, including:
  - Preparing well-documented data sets in Stata format
  - Preparing technical documentations highlighting the sampling design, data collection strategies and brief analyses
  - An In-depth study report including an executive report and detailed transcription of the in-depth interviews

- Infrastructure for hosting the data for public access

**Methodology**

- Participate in discussions on what data should be shared; how to share? What are the issues that need to be tackled when making the data available for public access?

- Develop a Data Management Plan

- Implement the activities sketched out by the Data Management Plan

***Project outputs***

Below is the list of key outputs:

- Two trips to IDRC in Ottawa
- Data Management Plan for Project 107145 "Strengthening the Economic Committee of the National Assembly in Vietnam"
- Implementation of the activities discussed in the Data Management Plan

In addition, the data sets collected by Project 107145 have been made publicly available at the following repository: https://goo.gl/eBj5mV

This is a folder for hosting the data temporarily while waiting for the official repository to be completely built at: rtopendata.org. Our ambition is that with a strong IT background, we will voluntarily build up and host a public repository for publicly available data sets.

***Problems and challenges***

There are a number of issues and challenges we have identified along the implementation of the project:

- Sharing data publicly is not a mindset for most researchers, ourselves included. Thus, an important task to do is to change the mindset that data need sharing
- Once researchers want to share, there are technical challenges that may prevent them from sharing properly. Sharing properly means we have to protect confidentiality and privacy of respondents; setting up the right infrastructure for sharing (server, repository, ...); developing necessary documentations for supporting data users in how to use the data sets; and having personnel to respond to inquiries by data users along the way.
- Executing a sharing protocol requires resources, which usually were not budgeted at the beginning

***Overall assessment and recommendations***

The project's implementation has been successfully and satisfactorily. It has met the objectives proposed in the proposal.

Appendix: Data Management Plan

# DATA MANAGEMENT PLAN

## STRENGTHENING THE CAPACITY OF THE ECONOMIC COMMITTEE OF THE NATIONAL ASSEMBLY IN INTEGRATING INCLUSIVE GROWTH IN MACROECONOMIC POLICY MAKING AND OVERSIGHT IN VIETNAM

#### ADMIN DETAILS

**Project Name:** My plan (Portage Template)
**Principal Investigator / Researcher:** Cameron Neylon
**Institution:** Portage

#### DATA COLLECTION

*WHAT TYPES OF DATA WILL YOU COLLECT, CREATE, LINK TO, ACQUIRE AND/OR RECORD?*

The project collects primary data from interviews with 760 SMEs in 12 provinces of Viet Nam. The sample was drawn from a stratified random procedure.

Throughout the 3-year period of the projects 4 rounds of the surveys have been conducted. The baseline surveys were conducted in the form of face-to-face interviews, with data entries being conducted with the rtSmartSurvey CAPI platform of Real-Time Analytics (RTA). The follow-up rounds are done on phones with data being entered to the electronic questionnaires while an interview is being conducted.

In short, the project generates data from primary data collections over 760 SMES. The data collections are rolled out into 4 rounds over the course of 3 years

*WHAT FILE FORMATS WILL YOUR DATA BE COLLECTED IN? WILL THESE FORMATS ALLOW FOR DATA RE-USE, SHARING AND LONG-TERM ACCESS TO THE DATA?*

Data are first stored in MySQL Database system of rtSmartSurvey platform. From there, the data are exported to Stata and CSV formats for analysis purposes.

For reuse, sharing and long-term access, Stata and CSV formats are highly relevant.

*WHAT CONVENTIONS AND PROCEDURES WILL YOU USE TO STRUCTURE, NAME AND VERSION-CONTROL YOUR FILES TO HELP YOU AND OTHERS BETTER UNDERSTAND HOW YOUR DATA ARE ORGANIZED?*

For activities under the project, we mostly work with data in Stata format. Stata data files has an embedded meta-data system, that includes variable labels and data set labels. The labels help explain what the variables are, when the dataset was created and (last) modified.

To help data analysts even further, we name the variables following the question numbering in the questionnaire.

**DOCUMENTATION AND METADATA**

*WHAT DOCUMENTATION WILL BE NEEDED FOR THE DATA TO BE READ AND INTERPRETED CORRECTLY IN THE FUTURE?*

The most useful documentation is the codebooks of the datasets in PDF format. The codebooks provide lots of meta data such as the total numbers of variables and observations; for each variable, it gives fundamental summary statistics and examples of data values.

The second documentation is technical notes such as sampling design note, data entry system and fieldwork plans.

The third documentation is a collection of descriptive analyses of the data sets.

*HOW WILL YOU MAKE SURE THAT DOCUMENTATION IS CREATED OR CAPTURED CONSISTENTLY THROUGHOUT YOUR PROJECT?*

There are three major mechanisms for ensuring consistency of the documentations throughout the project, including:

- Automating the creating of documentation whenever it's possible: we rely on rtSmartSurvey to generate the codebooks automatically
- Thorough reviewing of written reports and documentations. By being thorough I mean we get the papers reviewed by multiple staff in several rounds.
- Continuous revising: whenever there is a comment over the data that needs a revision, we go back to the dataset and conduct it. Then, an update would be added to the documentations.

*IF YOU ARE USING A METADATA STANDARD AND/OR TOOLS TO DOCUMENT AND DESCRIBE YOUR DATA, PLEASE LIST HERE.*

For the codebooks of the data sets we use the Stata in-built tools to translate from log files to PDF.

For other documentations, they are just reports/papers written in MS Office formats.

**STORAGE AND BACKUP**

*WHAT ARE THE ANTICIPATED STORAGE REQUIREMENTS FOR YOUR PROJECT, IN TERMS OF STORAGE SPACE (IN MEGABYTES, GIGABYTES, TERABYTES, ETC.) AND THE LENGTH OF TIME YOU WILL BE STORING IT?*

Since most data are in text format, all data storage is expected to be less than 100 Megabytes.
We envision that we will store it for at least 10 years.

*HOW AND WHERE WILL YOUR DATA BE STORED AND BACKED UP DURING YOUR RESEARCH PROJECT?*

We store the data sets in our staff's computers and a server operated by Real-Time Analytics.

*HOW WILL THE RESEARCH TEAM AND OTHER COLLABORATORS ACCESS, MODIFY, AND CONTRIBUTE DATA THROUGHOUT THE PROJECT?*

Currently, we share the data sets with the research team and other collaborator via dropbox sharing and emails.

**PRESERVATION**

*WHERE WILL YOU DEPOSIT YOUR DATA FOR LONG-TERM PRESERVATION AND ACCESS AT THE END OF YOUR RESEARCH PROJECT?*

Initially, we were considering developing an API (Application Program Interface) for accessing the data and hosting it at RTA's website at: www.rta.vn.
Nevertheless, after discussing with Prof. Neylon, we decide to host the data at a public Open Data Access service, such as:

- http://www.re3data.org:
- http://www.opendoar.org:

We will finalize the select of a repository after the wrap up meeting in December, 2016.

*INDICATE HOW YOU WILL ENSURE YOUR DATA IS PRESERVATION READY. CONSIDER PRESERVATION-FRIENDLY FILE FORMATS, ENSURING FILE INTEGRITY, ANONYMIZATION AND DE-IDENTIFICATION, INCLUSION OF SUPPORTING DOCUMENTATION.*

As described above, we think API-based protocol is best for this requirement.

**SHARING AND REUSE**

*WHAT DATA WILL YOU BE SHARING AND IN WHAT FORM? (E.G. RAW, PROCESSED, ANALYZED, FINAL).*

We will share the final data sets as open access and the raw data upon request.

*HAVE YOU CONSIDERED WHAT TYPE OF END-USER LICENSE TO INCLUDE WITH YOUR DATA?*

Not yet. We would be keen on learning relevant options and choose one that fits the best.

*WHAT STEPS WILL BE TAKEN TO HELP THE RESEARCH COMMUNITY KNOW THAT YOUR DATA EXISTS?*

There are several steps we can do, including:

- Mentioning the free access policy in the project's papers and reports
- Advertising the policy during workshop occasions
- Advertising the policy on websites of RTA and other project partners (CAF and ECNA)

**RESPONSIBILITIES AND RESOURCES**

*IDENTIFY WHO WILL BE RESPONSIBLE FOR MANAGING THIS PROJECT'S DATA DURING AND AFTER THE PROJECT AND THE MAJOR DATA MANAGEMENT TASKS FOR WHICH THEY WILL BE RESPONSIBLE.*

It shall be Real-Time Analytics as the institution and RTA's rtSolutions Team as staff who will be communicating on data access matters.

*HOW WILL RESPONSIBILITIES FOR MANAGING DATA ACTIVITIES BE HANDLED IF SUBSTANTIVE CHANGES HAPPEN IN THE PERSONNEL OVERSEEING THE PROJECT'S DATA, INCLUDING A CHANGE OF PRINCIPAL INVESTIGATOR?*

Since we work as a team we have staff backup plan on a daily basis. I don't expect major changes when a staff who is directly involved in the project leaves.

*WHAT RESOURCES WILL YOU REQUIRE TO IMPLEMENT YOUR DATA MANAGEMENT PLAN? WHAT DO YOU ESTIMATE THE OVERALL COST FOR DATA MANAGEMENT TO BE?*

The most resource consuming tasks are the development of the documentations, the development of API system and the continuous maintaining of staff who provides users with supports and feedback.

We will discuss with the team and develop a proposal for these tasks.

**ETHICS AND LEGAL COMPLIANCE**

*IF YOUR RESEARCH PROJECT INCLUDES SENSITIVE DATA, HOW WILL YOU ENSURE THAT IT IS SECURELY MANAGED AND ACCESSIBLE ONLY TO APPROVED MEMBERS OF THE PROJECT?*

We will trim all the identity information/variables from the data sets before we release to public users. In our data sets, these variables include company's name, tax code, phone number, address, and names of representatives. All observations and data points are then anonymous.

*IF APPLICABLE, WHAT STRATEGIES WILL YOU UNDERTAKE TO ADDRESS SECONDARY USES OF SENSITIVE DATA?*

The use of sensitive data (respondents' identity information such as names, addresses and telephones) is strictly within Real-Time Analytics for the purpose of following up with the respondents.

*HOW WILL YOU MANAGE LEGAL, ETHICAL, AND INTELLECTUAL PROPERTY ISSUES?*

We haven't thought much about this line of issues yet.