MANCHESTER 1824
The University of Manchester

Open Data Watch, Washington D.C: Access to Microdata (ATM) - Anonymization

Philippe Larsson, BSocSc (Hons) Politics and International Relations

Introduction

Open Data Watch (ODW) is an international non-profit, Non-governmental Organization (NGO) which specializes in helping National Statistical Offices (NSOs), governments and other organizations in their attempts to produce and manage official statistical data. Their efforts are centered around poorer, developing countries, but not exclusively. Their latest report, the **Open Data Inventory (ODIN) 2018/2019,** incluided 178 countries from around the world. Recently, however, ODW expanded their scope and started a new project, Access to Microdata (ATM), which aims to create a template of best practice for NSOs and governments to use when wanting to collect and publicize official statistical data. My work as a consultant ranged from note-taking at international meetings to cleaning up large datasets, but was mostly focused around researching anonymization methods for the ATM project.

Objectives

Overall, the objective of the Access to Microdata was to create a best practice protocol which would serve NSOs and governments wanting to better manage their official statistics - an important cause during a period in time that is sometimes referred to as the data revolution. Data is becoming increasingly available for organizations, governments and individuals on the internet, and Open Data Watch thus seeks to create means to better harness this data revolution.

On its own, the term 'open data' often raises questions around privacy issues. That is why a big part of creating this protocol was to ensure that **anonymization methods** are thoroughly researched and reformulated into a comprehensive checklist to ensure that the information collected on people is securely protected. When anonymization practices are maximized, chances of abuse, misuse and misrepresentation are minimized.

Method

In order to properly address the question of anonymization, the first step in my research was looking at the methodologies of a great many organizations. These included the **Demographic Health Surveys (DHS)**, the **UNICEF Multi Indicator Cluster Surveys (MICS)** and the **Living Standards Measurement Studies (LSMS)**, and served to get a feeling of what practices are already used by the more renowned survey oriented organizations.

The research was cross-referenced with existing academic literature on the topic, which tends to use coding and mathematical language. This made it hard to translate the litterature into exemplar practices to be added to the ATM protocol, but the overarching ideas confirmed and complimented the prior research, and some conclusions started to form.

The progress of the anonymization research was discussed at weekly briefings with the rest of the team. The ATM project is still in progress and includes many other themes, including best practices for ensuring **timeliness of data publicization**, evaluating **data accessibility**, and many more components that, if followed, will help in harnessing the data revolution.

Results and Conclusions

There are many ways to ensure that individuals remain anonymous, but two categories are essential to address, namely **Direct Identifiers** and **Attribute Identifiers** which are then **suppressed** or **generalized**.

Type of information	Method of anonymization	Identifying information
Direct Identifiers	Suppression	Names; addresses; GPS locations; phone numbers.
Attribute Identifiers	Generalization	Age may be presented in ranges; several identifiers may be collapsed into one (rare deceases categorized as "rare diseases" without disclosing the specific name); geographic coordinates of surveys may be displaced at a random distance and in a random direction.

Although this specific part of my internship did not include much quantitative research, it was the most significant and labor intensive part of it. Ensuring anonymity is very important when conducting surveys for publicization as it one of the key determinants for the quality of the survey. Not ensuring anonymity both endangers the participant in terms of online security and identity theft, and puts in jeopardy the organization responsible for survey. In some cases, prison sentences are the consquence of improper anonymization.

Key Skills Learnt

- Extensive knowledge of microdata portals (anonymization research)
- Knowledge on cybersecurity (anonymization research)
- Proficient user of Microsoft Excel (timeliness research)
- Intermediate user of APIs and CSV interface (accessibility research)
- Note-taking at meetings (overall)
- Establishing formal and informal work connections (overall)
- Engaging in roundtable discussions and international meetings (overall)

Produced at the Graphics Support Workshop, Media Services, The University of Manchester

OPEN DATA WATCH

Contact
philippe.larsson@student.manchester.ac.uk

Q-Step
A step-change in quantitative social science skills
Funded by the Nuffield Foundation, ESRC and HEFCE