# **Proof of Concept: collection, analysis and archiving of publicly available social media**

# Elaboration I Planning, and

# Elaboration II Testing and Revision

**Roslyn Walker Student Number**

This Proof of Concept proposes a workflow process for the collection, analysis, and archiving of publicly available data from online and social media platforms.

Both Elaboration I and Elaboration II plans are included in this documentation.

**Elaboration I** is a plan to identify a process for workflow for the Proof of Concept. The technical and non-technical considerations of workflow are in the scope of this Proof of Concept to address the ethical and legal framework for using personally identifiable information that forms the source data for this project.

The testing of this plan is set out in section two of this document, **Elaboration II**: Testing and revision. Elaboration II will be revised as testing for Elaboration I continues taking into account the outcome of the tests.

The aim of this plan is determine a set of technologies using a combination of tools, techniques and processes that work together to provide a solution to the problem of collecting source data from social media and archiving it (as defined in the scoping study for this project).

## Seven duties for the collection, analysis and archiving of publicly available social media

**1. Meet ethical research requirements**

Expected performance.

Collection of data meets research protocols.

This means compliance with the terms of service of social media platforms, specifically (ref).

The collection of personally identifiable information metadata will be minimised during collection

**2. Access source data online and from social media platforms**

Obtain Application Programming Interfaces (API)

API access Source data can be accessed: Non API methods.

Application Programming Interfaces (APIs) will be used in this project to enable technologies to communicate with each other.

**3. Store and retrieve source data**

For archival research it's important to build a pipeline around the management of the archival sources such as the images and text. Consider potential of a combination of application in existing platforms and systems with open source tools including Zotero as a bibliography software and Tropy as a research photograph tool.

**4. Analyse source data**

Format of data. CSV.

thematic tagging of content

Display of content as a virtual exhibition

**5. Data governance maintained through the solution workflow**

Throughout the workflow process the solution has a duty to include a data governance framework that provides checks-and-balances to maintaining the integrity of the data through the workflow. The governance framework will use the five knows principle:

1. Do you know the value of your data?

2. Do you know who has access to your data?

3. Do you know where your data is?

4. Do you know who is protecting your data?

5. Do you know how well your data is protected?

# Elaboration II Testing and Revision

## 1. Ethical (Testing duty 1)

Test. Research copyright and download restrictions of commercial entities by looking at their terms of service and identifying what can or can't be done with the data.

## 2. Source data test (Testing duty 2)

Test. Extract data from Twitter. Successful in initial test to retrieve my own personal data files from Twitter. Process

Test. Test whether an API be retrieved from Twitter. Test by contacting Twitter and

Test. Automation process of going to a platform retrieving data and displaying it.

## 3. Storage and retrieval (Testing duty 3)

Test. Shell scripting to convert source data to machine readable text for storage/archiving.

Test. Store the data and retrieve in manipulated form.

## 4. Analysis (Testing duty 4)

Test. Whether data can be manipulated and rearranged.

Test. Locate dataset tools that can help me to interact with the data to enable formatting and output.

Test. Shell scripting to do a thematic analysis of the text.

## 5. Pipeline connection (Testing duty 3)

Test. Workflow between sourcing, storing and retrieving data.

Test. Use of and connection of proposed tools and techniques identified through testing to provide a solution to the problem of collecting source data from social media, storing it and manipulating it.

Test. Existing platforms and systems as possible solution.

Test: The steps in the solution and whether the test plan and scope of project allows sufficient time to learn, use and connect the proposed tools and techniques in a 30 day timeframe.

## 6. Data governance (Testing duty)

Test. Automation of backups and manual testing procedure.

Test. Capture process of metadata with file creations.

Test. Password storage and retrieval on applications with access to data. Test application of LastPass in the solution for data governance.

**References:**

Connolly, B. (2015). CISO Telstra’s ‘five knows of cyber security’ at www.cio.com.au/article/583438/telstra-five-knows-cyber-security

**Elaboration one.**

Task:

(1) Use the step-by-step breakdown of your problems and solutions from the previous week (such as programming languages, software libraries, APIs (Application programming interfaces, the means by which one program talks to another program).

(2) Write in LaTeX

(3) Testing the steps: Can you see yourself learning enough to use and connect these tools and techniques in the next month?

Collection of data to meet research protocols.

**Data Governance**