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

# Elaboration II Testing and Revision

**Roslyn Walker Student Number 45149631**

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

The aim of this plan is to test and revise a set of technologies 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).

The seven duties for the collection, analysis and archiving of publicly available social media are outlined in the Elaboration I plan. These duties will be tested and revised as set out in this document. Elaboration II will be revised as testing for Elaboration I continues taking into account the outcome of the tests.

## 1. Ethical (Testing duty 1)

Test. Research copyright and download restrictions of commercial entities by looking at their terms of service and identify what can or can't be done with the data. View user agreement for Twitter. Identify relevant sections. Alternatively utlise public interface of curating a collection. Locate further detail on this way forward in real time: https://developer.twitter.com/en/docs/tweets/curate-a-collection/api-reference or access to 7 days / 30 days of tweets through search.

## 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. For further elaboration on this test API results through shell scripting.

Next steps 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. Store the data and retrieve in manipulated form. Initial shell scripting completed through Data Carpentry exercises. Next steps to tailor to data source files on API Tweet files.

## 4. Analysis (Testing duty 4)

Test. Whether data can be manipulated and rearranged. Next steps locate dataset tools that can help me to interact with the data to enable formatting of output and thematic analysis of the text.

## 5. Pipeline connection (Testing duty 3)

This test step is out of scope for this work and will exceed initial 30 days of testing solutions. Workflow between sourcing, storing and retrieving data. 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.

If sufficient time to learn, use and connect the proposed tools and techniques this will be covered in scope.

## 6. Data governance (Testing duty 5 )

Test. Automation of backups and manual testing procedure. Capture process of metadata with file creations. Password storage and retrieval on applications with access to data. Tested application of LastPass in the solution for data governance. Tweet password applied as initial test for application integration through password access.