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

# Elaboration I Planning

**Roslyn Walker Student Number 45149631**

*Task: Using the step-by-step breakdown of your problems / solutions developed last week, identify technologies (programming languages, software libraries, APIs (Application programming interfaces, the means by which one program talks to another program) and other components of a modern data collection or processing workflow) that could accomplish each step. You may want to specify more than one possible technology.*

*This document can be written in any format (though we encourage further practice in LaTeX) and should be as long as it needs to be. If you find yourself testing too many steps or technologies which depend on each other, you may wish to adjust your scope to be more specific. The judgment call for “too many” is “Can you see yourself learning enough to use and connect these tools and techniques in the next month?”*

*Alternatives found for components which do not prove suitable during elaboration. All components identified could be credibly combined together to solve the problem articulated in the scoping document and the steps necessary to achieve this goal have been identified.*

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

**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.

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

**1. Meet ethical research requirements**

With reference to the Guide to data analytics and the Australian Privacy Principles from the Office of the Australian Information Commission--the collection of data from Twitter must comply with Australian laws. In addition as research is conducted with Macquarie University for learning and research the process must comply with ethical research protocols. This also means the compliance with the terms of service of social media platforms.

One approach is to investigate the minimisation of the collection of personally identifiable information metadata will be minimised. The use of ‘all the data’ for ‘unknown purposes’ exposes the University to privacy compliance risks. This requires the limiting of the collection of personal information that I am only collecting information which is reasonably necessary to pursue the legitimate functions and activities of this project.

With respect to sensitive information the creation of new data sets may lead to the creation and collection of sensitive information may be generated from inferred or derived data. Where personal information is created which the organisation is not able to collect under APP3, it may need to be de-identified or destroyed. See [Collecting Personal Information](https://www.oaic.gov.au/privacy/guidance-and-advice/guide-to-data-analytics-and-the-australian-privacy-principles/#s2-2-collection-of-personal-information-app-3) in Part Two." (OAIC 2019).

An alternative to this approach is to use the data sets generated through my own Twitter feed for the Proof of Concept and investigate whether this creates compliance issues with regard to ethical protocols or privacy compliance.

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

Twitter feed data is obtain through an Application Programming Interfaces (API). Twitter is able to provide access to API through a developer account. Atlernatively this is provided for use on my own data held by Twitter. As an alternative path this Proof of Concept could explore non API methods. The Application Programming Interfaces (APIs) could 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. The Proof of Concept could consider the potential of using a combination of applications 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**

Explore file formats from source data on Twitter with potential thematic tagging of content and 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?

Data governance software includes

**References:**

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

OAIC 2019, the Guide to data analytics and the Australian Privacy Principles https://www.oaic.gov.au/privacy/guidance-and-advice/guide-to-data-analytics-and-the-australian-privacy-principles/)