Curtin University

Research Data Management Plan

Piloting Data Collection and Management for a Usage Data Trust

Principal Investigator	Lucy Montgomery
Data Management Plan Edited by	Alkim Ozaygen
Modified Date	20/05/2020
Data Management Plan ID	MONTGA-HU07600
Faculty	Humanities

1 Research Project Details

1.1 Research project title

Piloting Data Collection and Management for a Usage Data Trust

1.2 Research project summary

This research project is focussed on developing the tools, software and workflows needed to gather, process and synthesise usage data related to OA scholarly books.

The project will feed into a larger Andrew W. Mellon Foundation funded initiative titled: Developing a Data Trust for Open Access Ebook Usage. Full details of the larger initiative are available here: https://digital.library.unt.edu/ark:/67531/metadc1596980/

1.3 Keywords

big data, monographs, open access

2 Research Project Data Details

2.1 Research project data summary

The project will comply with licensing and contractual agreements with publishers, repositories, and platforms. Publishers includes organisations that publish scholarly monographs including university presses, small independent presses, and commercial organisations. The partner publishers will be identified in an early phase of the project. Repositories includes institutional repositories that host books (such as Curtins eResearch Repository) and large general archives such as the Internet Archive. Platforms includes specialist sites that make scholarly content available on behalf of publishers including OAPEN, JSTOR and MUSE.

This project involves the following types of data:

Usage data - Usage data relates to how and when an OA scholarly book has been accessed via the web. It includes usage logs and web-view data such as Piwik or Google Analytics.

- This usage data is already being generated/collected by repositories, publishers, and websites.
- For example this usage data may be generated directly via the publisher's own website, or via platforms that distribute OA scholarly content, such as OAPEN, JSTOR or MUSE.
- Usage data is generally commercially sensitive data. It may include information such as country, IP address (anonymised by removing the last byte), the date the access occurred, the form of usage (download, web page view), and the content accessed (book, chapter, abstract, book presentation page).

Additional indicators of use

- Additional, public, sources of data may also provide indications of use for OA scholarly books.

- Publicly available ‘altmetrics’ will be gathered via public APIs or, where possible, direct data dump, from services that include Twitter, Crossref, Worldcat and Mendeley.

- This data may include information on public user identity, public name, date, type of value, entered value (comment, mentions, etc.), book title or book identifier.

2.2 Will the data be identifiable

- Individually identifiable — the identity of a specific individual can reasonably be ascertained (e.g. name)
- Non-identifiable — data which has never been labelled with individual identifiers

2.3 Will biospecimens or human participant information be sent overseas?

Yes

2.3.1 Indicate where it will be sent, in what format and how you have complied with legislation regarding transfer of data

Anonymised usage logs provided by partnering platforms and publishers; as well as publicly available social media data; will be shared with project partners in aggregate form, as well as via interactive visualisation dashboards. Project partners are located in the US, Europe, South Africa, and Australia.

Project data will also be stored in Google BigQuery, and hosted on the Google Cloud Platform.

2.4 Will novel information about controlled goods or technologies on the Defence and Strategic Goods List (DSGL) be sent overseas?

No

2.4.1 Indicate where it will be sent, in what format and how you have complied with legislation regarding transfer of data

Not Applicable

2.5 Data organisation and structure

Data from partners will be stored in the form of individual datasets created for each partner data source (i.e. repository, publisher, platform). Within each dataset we will create one or more separate tables for each type of data collected. All data will be time-stamped so that it will be possible to know when an individual data set was generated. and information on metadata. The metadata will contain information such as the entry id, title name, identifier (DOI, ISBN, URL, tweet ID), country name, date-time, type of data, type of value, value, date stamp, amongst others. End-user documentations will be for users who are going to use dashboards of the data trust.

As this data will primarily be stored in a relationship database, schemas will be created and documented for each table. Additionally, these tables can be exported into CSV and JSON formats for long term storage. Associated with the explicit schemas, which specify the syntactic structure of the datasets, we will produce human readable descripting to save key semantic and provance information about these datasets.

In addition to commentary on the data itself, documentation will include information on infrastructure, installation, and organisation of the code that was used to gather and process these datasets.

3 Research Project Data Storage, Retention and Dissemination Details

3.1 Storage arrangements

All the data will be stored in the Google Cloud Platform which is within the scope of allowed data storage solutions as advised by the E-Infrastructure Working Group. The data will primarily be stored inside a relational database. The database tools available within Google cloud can be fully managed, with a pricing model that is based on storage volume and usage, as opposed to any one-time or recurring licensing fees. This means that retention of the data for required periods can be maintained at a low cost.

Additionally, snapshots of this data will be made in both CSV and JSON formats that will be saved in Cloud Storage, an object-based storage solution designed for very high levels of availability and durability.

Logically, all Google Cloud storage solutions are global in scope, having the same interface and security model. However, the larger Data Trust project includes funding for a legal consultant, who will provide advice on issues relating to the physical location of the data. Our team will adhere to this advice in order to ensure that data is handled and stored in accordance with relevant legislation.

The Curtin R-Drive is not a suitable storage or archiving solution for this data due both to its form and size.

3.2 Estimated data storage volume

Estimating the data storage requirements is premature currently. This will primarily be driven by the number of partners and their willingness to share their data. The types of data that might be made available to use are unlikely to cause issues here though. We are not working with video, audio or volumetric data. There will also be opportunities to optimize the type of data we store and keep to keep storage volumes low. The cloud model we are using places no practical constraints on the amount of data we can store, and does not force us to have accurate estimates before setting off.

The data volume will increase over time but is expected to be on the order of terabytes (TB) in total.

3.3 Safeguarding measures

The system (including the data and the code we run) will be implemented on top of a commercial cloud platform (Google Cloud Platform) utilising a defense in-depth strategy. - All data stored at rest or in transit in our system will be encrypted. - Role-based access control (using the Cloud IAM service) will be in force at all times. Meaning that you must not only have a user account to the system, but for that user to have the required permissions to access or modify any piece of data or the systems that process this data. Access Groups will be used to manage privileges to access or query data, with individuals added or removed from these groups. This separates out the task of carefully defining permissions, from the management of who is granted these permissions. - The data centers running the underlying systems are run in physically secure locations and this hardware the the software layers above it comes with a range of independently audited compliances: https://cloud.google.com/security/compliance - The two primary mechanisms we will provide partners with access to either the raw data or dashboards will be through BigQuery and Data Studio. Data Studio is built directly on BigQuery and both have access control methods that can limit access to the whole dataset, individual tables or even specific rows and columns within the table. This is control via the access groups and Cloud IAM as mentioned above. - All data stored within BigQuery (along with other hosted services) is replicated across multiple availability zones (physical building with independent electrical systems for example) or if we choose to pay slightly more, across geographical regions (say the east and west coasts of the USA). We will save exported copies of our databases in CSV and JSON formats with Cloud Storage with also comes with similar levels of redundancy.

3.4 Retention requirements

7 years (All other research with outcomes that are classed as Minor)

3.5 Collaboration

This project involves stakeholders such as publishers, repositories and Internet platforms, all of which will share their data with the data trust we are building. Data belonging to partners will be collected automatically using drivers or manually by the stakeholders.

These partners will have access to the whole of the data that they contribute on the data trust. The purpose of the project is to define how aggregate and derived data products might be shared in a trusted fashion between the partners (including publishers, repositories, and platforms) and what arrangements are required to ensure this trust. For example

a publisher dashboard might contain information on the usage for that publishers books and a comparison to an aggregate benchmark of usage from other publishers. Alternately, some publishers or platforms may choose to share aggregate and de-identified usage data freely within the project partners, or publicly.

Data access policies will be designed collaboratively through the project with all stakeholders involved in this process. This policy will define how to preserve commercially sensitive information, while collectively sharing the benefit of a wider community understanding.

3.6 Data dissemination

During the project a data sharing and dissemination policy will be designed. After the design is developed and shared with stakeholders, the data collection will start. All the data acquired from a partner will be shared only with the partner's authorization.

All personally identifiable data will be anonymised before it reaches our systems. If the partner lacks the in-house technical skills to de-identify data before it is provided to us, we will take steps to work with them to de-identify data and remove the identifiable information. The project will create at least three prototype dashboards for different types of partners. As previously mentioned, this will be built using Data Studio, with the data itself hosted within BigQuery. Beyond this funded project, these dashboard might be developed further using a different platform, however this work is beyond the scope of this document.

3.7 Embargo period

There is no planned embargo period for the collected data. However, the partners may choose to make public their closed access data after a period of time. In that case, an embargo period policy can be designed.