Data Management Plan Exemplar #2: Digital Humanities and Secondary Data

Historical Canadian Census Data

Abstract

Text-based data collected from Statistics Canada was used to create a union list of born-digital products from the Canadian Census of Population, starting with the 1961 Census. This union list indicates where the census files are located in Canada (for example, the University of Toronto Data Library) and what they contain. The data is stored in a database and accessible through an online search engine (see: Search for Aggregated Data Files from Canadian Censuses URL: http://mdc.lib.uwo.ca/census/pubsearch.htm)

Principal investigators:

Vincent Gray (Western University)

Alexandra Cooper (Queen's University)

Administrative Details

Project Name: Historical Canadian Census Data

Principal Investigator / Researcher:

Vincent Gray (Western University); Alexandra Cooper (Queen's University)

Project Data Contact:

Vince: vince@uwo.ca; Alex: coopera@queensu.ca

Description:

This project will create a bilingual union list of born-digital products from the Canadian Census of Population, starting with the 1961 Census. The database will be accessible through an online search page to allow users to identify appropriate files to use at the required level of geography. It will also have an online editing function to allow institutions to supplement the database with local holdings information.

Data Collection

What types of data will you collect, create, link to, acquire and/or record?

Text-based data are collected from Statistics Canada, falling under the Statistics Canada Open License.

What file formats will your data be collected in? Will these formats allow for data re-use, sharing and long-term access to the data?

The data will be collected in an Inmagic v.15 database from which non-proprietary file-formats, including ASCII, HTML and/or XHTML text files, may be extracted. Inmagic is a content management system, a type of database used to manage large amounts of content, such as documents, images, and more.

What conventions and procedures will you use to structure, name and version- control your files to help you and others better understand how your data are organized?

As records are created and/or edited, they will be time-stamped with changes. The name of the person making the changes is also requested. A log file is kept to record information about database changes.

Documentation and Metadata

What documentation will be needed for the data to be read and interpreted correctly in the future?

Four different pieces of documentation will be needed.

1. A description of the project, which will include a description of the process undertaken to identify the various historical census data files;
2. A description of the field structure in Inmagic (e.g., whether a field is required, uses a controlled vocabulary, is repeatable, etc.);
3. The data entry instructions to be followed in populating the database; and

4. Lists showing the possible values of the various controlled-vocabulary fields, and of the substitution lists used to automatically translate English to French text and vice versa within the paired controlled-vocabulary fields.

How will you make sure that documentation is created or captured consistently throughout your project?

1. A working document will be created and revised as new content is added to the
1. database.
2. The field structure is stored within Inmagic and may be printed out as a text file.
3. Data entry instructions are included on the web-based data entry form within Inmagic and may be printed out as a text file.
4. These lists are stored within Inmagic and may be printed out as text files.

If you are using a metadata standard and/or tools to document and describe your data, please list here.

It would be possible to enhance the records by including complete documentation as per the Data Documentation Initiative. However, that is not yet within the purview of this project.

Storage and Backup

What are the anticipated storage requirements for your project, in terms of storage space (in megabytes, gigabytes, terabytes, etc.) and the length of time you will be storing it?

Currently, the Inmagic database has approximately 640 records and occupies approximately 4.7 MB across 11 proprietary format files. Anticipating that the database will grow in size to approximately 10,000 records, it might be anticipated to grow to no more than 100 MB. Additionally, regular file dumps in ASCII format will be performed, to ensure that the contents of the database will be transportable to other database systems or used by other interfaces: the records in the database currently occupy about 2K per record in delimited format, and compress from 1,220 K to 94 K. At their largest, each backup file might be expected to require 1.5 MB in compressed format.

There is no anticipated timeline on retiring this project: space requirements would continue to grow as the project adds other census years. These additional storage requirements have been accounted for by IT Service, therefore ensuring that future space needs will be met.

How and where will your data be stored and backed up during your research project?

The data are stored on Western University servers which are connected to automated backup systems.

How will the research team and other collaborators access, modify, and contribute data throughout the project?

Access to the database will be either directly though the Inmagic interface for batch loading (V. Gray), or through the web-based interface (A. Cooper and other contributors). Access to editing the database will be restricted. Contributors will be given a user ID and password to allow for editing.

Preservation

Where will you deposit your data for long-term preservation and access at the end of your research project?

The project does not have a foreseeable end date. An ASCII delimited (and potentially a XHTML) version of the database will be created and could be stored on Scholars Portal Dataverse, a data repository which assigns DOIs to datasets, and supports preservation, discovery, citations, and data usage metrics. However, a consultation with the University's Research Data Management Librarian will help identify other possible repository options for our research data.

Indicate how you will ensure your data is preservation ready. Consider preservation-friendly file formats, ensuring file integrity, anonymization and de-identification, inclusion of supporting documentation.

Preservation format copies of the database will be stored in ASCII delimited format. As new versions are created, they will be compared to previous versions to ensure that the previous versions contain the same data for unmodified records as the new. The required documentation files will be saved on the preservation site along with the delimited files.

Sharing and Reuse

What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final).

The raw data (i.e., bibliographic and holdings information) will be shared.

Have you considered what type of end-user license to include with your data?

Creative Commons Attribution-Sharealike 4.0.

What steps will be taken to help the research community know that your data exists?

To date, presentations have been made at regional Data Liberation Initiative training sessions for the Ontario, Western Canada, and Atlantic regions, and to Statistics Canada (Data Rescue and Recovery Update URL: https://cudo.carleton.ca/dli-training/4075). It is hoped that a presentation to the Quebec region will also be possible. Finally, an article may be written and submitted to Statistics Canada for inclusion in its DLI Newsletter and/or to an academic library journal to highlight the existence of the tool. Depending on the repository that the data is deposited in, there may be additional resources to notify the community. If deposited in Dataverse, a persistent digital object identifier (DOI) will be minted for the dataset providing a persistent identifier and improving chances of discoverability.

Responsibilities and Resources

Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible.

Vincent Gray: batch uploading of records in French and English, editing records, performing routine maintenance on the Inmagic database; extracting ASCII files; creating documentation; migration to preservation platform

Alexandra Cooper: creating and editing records; creating documentation; migration to preservation platform

How will responsibilities for managing data activities be handled if substantive changes happen in the personnel overseeing the project's data, including a change of Principal Investigator?

Automated processes will be established to regularly extract preservation files to disk that can be uploaded.

The Ontario Data Community (part of OCUL) will oversee the project if/when there is a change in Principal Investigators.

What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?

Storage space on a web-enabled server. The entire project would fit onto a 4-GB USB key with space to spare or may be written at intervals to DVD. Minimal long-term costs would be expected as long as Western maintains a web based Inmagic service: should this change, a new platform would need to be selected and created.

Ethics and Legal Compliance

If your research project includes sensitive data, how will you ensure that it is securely managed and accessible only to approved members of the project?

No sensitive data are included in this project: it will consist solely of pointers to files of data which were released for public use by Statistics Canada.

If applicable, what strategies will you undertake to address secondary uses of sensitive data?

Not applicable.

How will you manage legal, ethical, and intellectual property issues?

The data files recorded in the database are in the public domain, subject to licensing: the database itself will not contain these data.

This document was generated by DMP Assistant (https://assistant.portagenetwork.ca)