

Plan de gestion des données modèle #2

Sciences humaines numériques et données secondaires

Données historiques du recensement canadien

Résumé

Des données textuelles de Statistique Canada ont été recueillies afin de créer un catalogue collectif de produits numériques du Recensement canadien de population, à partir du recensement de 1961. Ce catalogue collectif indique l'emplacement au Canada des fichiers du recensement (par exemple, la Bibliothèque de données de l'Université de Toronto) ainsi que leur contenu. Ces données sont stockées dans une base de données accessibles par l'entremise d'un moteur de recherche en ligne. (voir : Rechercher des fichiers de données agrégées des recensements canadiens : <http://mdc.lib.uwo.ca/census/pubsearchf.htm>)

Chercheuse principale :

Vincent Gray (Western University)

Alexandra Cooper (Queen's University)

Détails administratifs

Nom du projet : Données historiques du recensement canadien

Chercheur principal/intendante des données de recherche :

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

Responsable des données du projet :

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

Description:

Ce projet permettra de créer un catalogue collectif bilingue des produits numériques d'origine du Recensement de la population du Canada, à partir du Recensement de 1961. La base de données sera accessible au moyen d'une page de recherche en ligne qui permettra aux utilisateurs de déterminer les fichiers appropriés à utiliser au niveau géographique requis. Elle comportera également une fonction de modification en ligne qui permettra aux établissements de compléter la base de données par des renseignements sur les fonds locaux.

Plan de gestion des données modèle #2 Sciences humaines numériques et données secondaires Données historiques du recensement canadien

Collecte des données

Quels types de données recueillerez-vous, créerez-vous, couplerez-vous, acquérez-vous ou consignerez-vous?

Les données textuelles sont recueillies auprès de Statistique Canada et relèvent de la licence ouverte de Statistique Canada.

Selon quels formats de fichiers les données seront-elles recueillies? Ces formats permettront-ils la réutilisation, le partage et l'accès à long terme aux données?

Les données seront recueillies dans une base de données Inmagic v.15 à partir de laquelle des formats de fichiers non exclusifs, y compris les fichiers ASCII, HTML ou XHTML, pourront être extraits. Inmagic est un système de gestion de contenu, c'est-à-dire un type de base de données servant à gérer de grandes quantités de contenu, comme des documents, des images et plus encore.

Quelles conventions et procédures utiliserez-vous pour structurer et nommer vos fichiers, et pour contrôler leur version, afin de vous aider, ainsi que les autres, à mieux comprendre comment vos données sont organisées?

À mesure que les enregistrements sont créés ou modifiés, ils seront horodatés avec leurs modifications. Le nom de la personne qui apporte les modifications est également demandé. Un fichier journal est conservé pour consigner l'information sur les modifications apportées à la base de données.

Documentation et métadonnées

Quels seront les documents nécessaires pour que les données soient lues et interprétées correctement à l'avenir?

Il faudra quatre documents différents.

1. Une description du projet, qui comprendra une description du processus entrepris pour déterminer les divers fichiers de données historiques du recensement;

Plan de gestion des données modèle #2 Sciences humaines numériques et données secondaires

Données historiques du recensement canadien

2. Une description de la structure de champ dans Inmagic (p. ex., si un champ est requis, utilise un vocabulaire contrôlé, est reproductible, etc.);
3. Les instructions de saisie des données à suivre pour remplir la base de données;
4. Les listes montrant les valeurs possibles des divers champs de vocabulaire contrôlé et les listes de substitution utilisées pour traduire automatiquement l'anglais en français et vice versa dans les champs de vocabulaire contrôlé appariés.

Comment vous assurerez-vous que la documentation est créée ou saisie de façon uniforme tout au long de votre projet?

1. Un document de travail sera créé et révisé à mesure que le nouveau contenu sera ajouté à la base de données
2. La structure de champ est stockée dans Inmagic et peut être imprimée sous forme de fichier texte.
3. Les instructions de saisie des données sont incluses dans le formulaire de saisie des données sur le Web dans Inmagic et peuvent être imprimées sous forme de fichier texte.
4. Ces listes sont stockées dans Inmagic et peuvent être imprimées sous forme de fichiers texte.

Si vous utilisez une norme ou des outils de métadonnées pour documenter et décrire vos données, veuillez les énumérer ici.

Il serait possible d'améliorer les notices en incluant la documentation complète conformément à Data Documentation Initiative. Toutefois, cela ne fait pas encore partie du projet.

Stockage et sauvegarde

Quelles sont les exigences prévues en matière de stockage pour votre projet, en ce qui concerne l'espace de stockage (en mégaoctets, gigaoctets, téraoctets, etc.) et la durée de stockage?

À l'heure actuelle, la base de données Inmagic compte environ 640 notices et occupe environ 4,7 Mo dans 11 fichiers en format exclusif. On prévoit que la base de données atteindra environ 10 000 notices et qu'elle ne dépassera pas 100 Mo. Par ailleurs, il y aura des vidages de fichiers périodiques en format ASCII pour faire en sorte que le contenu de la base de données soit transférable à d'autres systèmes de base de données ou qu'il soit utilisé par d'autres interfaces : les notices dans la base de données occupent actuellement environ 2 Ko par notice

Plan de gestion des données modèle #2 Sciences humaines numériques et données secondaires

Données historiques du recensement canadien

en format délimité, et peuvent être comprimés de 1 220 Ko à 94 Ko. Les plus gros fichiers de sauvegarde pourraient représenter 1,5 Mo en format comprimé.

Il n'y a pas d'échéancier prévu pour le retrait de ce projet : les exigences en matière d'espace continueront de croître à mesure que le projet ajoutera d'autres années de recensement. Ces exigences en matière de stockage supplémentaires ont été prises en compte par les Services de technologie de l'information, ce qui garantit que les besoins d'espace à venir seront satisfaits.

Comment et où vos données seront-elles stockées et sauvegardées pendant votre projet de recherche?

Les données sont stockées sur des serveurs de la Western University, lesquels sont connectés à des systèmes de sauvegarde automatisés.

Comment l'équipe de recherche et les autres collaborateurs pourront-ils accéder aux données, les modifier et y contribuer tout au long du projet?

L'accès à la base de données se fera soit directement par l'interface Inmagic pour le chargement par lots (V. Gray), soit par l'interface Web (A. Cooper et d'autres contributeurs). L'accès à la modification de la base de données sera restreint. Les contributeurs recevront un code d'utilisateur et un mot de passe pour apporter des modifications.

Conservation

Où déposerez-vous vos données pour la conservation à long terme et l'accès à la fin de votre projet de recherche?

Le projet n'a pas de date de fin prévisible. Une version ASCII délimitée (et potentiellement une version XHTML) de la base de données sera créée et pourrait être stockée dans [Scholars Portal Dataverse](#), un dépôt de données qui attribue des DOI aux ensembles de données et permet la conservation, la découverte, la citation et les statistiques d'utilisation des données. Toutefois, une consultation avec le bibliothécaire en gestion des données de recherche de l'université aidera à déterminer d'autres options de dépôt possibles pour nos données de recherche.

Plan de gestion des données modèle #2 Sciences humaines numériques et données secondaires

Données historiques du recensement canadien

Indiquez comment vous vous assurerez que vos données sont prêtes à être conservées. Envisagez des formats de fichiers favorables à la conservation, en assurant l'intégrité des fichiers, l'anonymisation et la dépersonnalisation, et l'inclusion des documents justificatifs.

Des copies de la base de données en format de préservation seront stockées en format ASCII délimité. À mesure que de nouvelles versions sont créées, elles seront comparées aux versions précédentes afin de s'assurer que les versions précédentes contiennent les mêmes données pour les notices non modifiées que pour les nouvelles. Les fichiers de documentation requis seront sauvegardés sur le site de préservation avec les fichiers délimités.

Partage et réutilisation

Quelles données partagerez-vous et sous quelle forme (p. ex., brute, transformée, analysée, définitive)?

Les données brutes (c.-à-d. renseignements bibliographiques et sur les fonds de données) seront partagées.

Avez-vous envisagé le type de licence d'utilisateur final à inclure dans vos données?

Attribution-Partage dans les mêmes conditions 4.0.

Quelles mesures seront prises pour aider à faire connaître l'existence de vos données dans le milieu de la recherche?

À date, des présentations ont été faites aux séances de formation régionales de l'Initiative de démocratisation des données (IDD) pour les régions de l'Ontario, de l'Ouest canadien et de l'Atlantique, et à Statistique Canada (<https://cudo.carleton.ca/dli-training/4075>). On espère qu'une présentation pour la région du Québec sera également possible. Enfin, il se peut qu'un article soit soumis à Statistique Canada pour inclusion dans son bulletin de l'IDD ou à un périodique en sciences de l'information, afin de souligner l'existence de l'outil. Selon le dépôt dans lequel sont déposées les données, il pourrait y avoir des ressources supplémentaires pour aviser la collectivité. Si les données sont déposées dans Dataverse, un identifiant numérique

Plan de gestion des données modèle #2 Sciences humaines numériques et données secondaires

Données historiques du recensement canadien

d'objet sera produit pour l'ensemble de données, ce qui fournira un identifiant pérenne et améliorera les chances de découverte.

Responsabilités et ressources

Déterminez qui sera responsable de la gestion des données de ce projet pendant et après le projet, et les principales tâches de gestion des données dont cette personne sera responsable.

Vincent Gray :

- téléchargement par lots des notices en français et en anglais, modification des notices, entretien de routine de la base de données Inmagic; extraction de fichiers ASCII; création de la documentation; migration vers la plateforme de préservation

Alexandra Cooper :

- création et modification des notices; création de la documentation; migration vers la plateforme de préservation

Comment les responsabilités relatives à la gestion des activités de données seront-elles gérées si des changements importants surviennent dans le personnel qui supervise les données du projet, y compris un changement de chercheur principal?

Des processus automatisés seront mis en place pour extraire régulièrement des fichiers de préservation sur disque pour pouvoir les télécharger.

La communauté des utilisateurs de données de l'Ontario (qui fait partie du Conseil des bibliothèques universitaires de l'Ontario) supervisera le projet s'il y a un changement au chapitre des chercheurs principaux.

De quelles ressources aurez-vous besoin pour mettre en œuvre votre plan de gestion des données? À combien estimez-vous le coût global de la gestion des données?

Espace de stockage sur un serveur Web. L'ensemble du projet peut être sauvegardé sur une clé USB de 4 Go, ou peut être sauvegardé à intervalles réguliers sur DVD. On s'attendrait à des

Plan de gestion des données modèle #2 Sciences humaines numériques et données secondaires

Données historiques du recensement canadien

coûts minimes à long terme tant que la Western University maintiendrait un service Web Inmagic; si cela devait changer, il faudrait sélectionner et créer une nouvelle plateforme.

Éthique et conformité juridique

Si votre projet de recherche comprend des données de nature délicate, comment veillerez-vous à ce qu'elles soient gérées de façon sécuritaire et accessible uniquement aux membres approuvés du projet?

Ce projet ne comprend pas de données sensibles; il s'agira uniquement de pointeurs vers des fichiers de données qui ont été diffusés au grand public par Statistique Canada.

Le cas échéant, quelles stratégies entreprendrez-vous pour aborder les utilisations secondaires des données sensibles?

Sans objet.

Comment allez-vous gérer les questions juridiques, éthiques et de propriété intellectuelle?

Les fichiers de données enregistrés dans la base de données sont du domaine public, sous réserve de licences; la base de données elle-même ne contiendra pas ces données.

Ce document a été produit par l'Assistant PGD.



Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
Portage Network | portage@carl-abrc.ca | portagenetwork.ca