Borealis to House Sensitive Metadata (by Thierry)

Introduction

Borealis is the Canadian Dataverse Repository. It holds data from most public Canadian universities in their own repositories called “Dataverses”. In essence, Borealis is the repository of those institutional repositories. It facilitates access to data - some of which is restricted and requires owner authorization to access the data itself. That being said, the metadata about these restricted datasets allow for their findability.

It logically follows from this that, given our goals of making restricted data findable with heavily anonymized metadata, borealis could be a central tool in our on-going endeavor. Using Borealis also has the added benefit of linking a persistent identifier to the dataset, and of being automatically converted into a standardized DDI XML file. Of course, there are certain qualms about using a national repository such as Borealis. Being a national service, Borealis data is often scraped by foreign academic colleagues in both peer and “near-peer” nations. Given the sensitivity of this data, how “findable” we want this data to be beyond the confounds of Canada remains a point of contention.

Nevertheless, there is reason to believe that this can be done safely and effectively to preserve the knowledge of these dataset's existence and allow for them to be traced back to a persistent identifier. The present document assesses the feasibility and benefits of using Borealis as a haven for sensitive data metadata.

Creation of the CRDCN Dataverse

Before even talking about how to safely upload metadata pertaining to sensitive datasets, it is important to have an idea of how this would be done in the first place. The first step would be for the CRDCN to set up a Dataverse of their own within Borealis. it would require discussions with the Borealis team as to where this Dataverse would be set up. For instance, not all Dataverses are immediately visible on the main Borealis page; for the most part, only academic institutions are advertised on the main page. However, there are other repositories that are “hidden” (such as the Dataverse for Odesi, the Social Science data repository operated by Scholars Portal).

The CRDCN has two main options:

1. Create their own Dataverse (not my go to)
2. Odesi Dataverse (whose explicit purpose is to house public access Statistics Canada data)

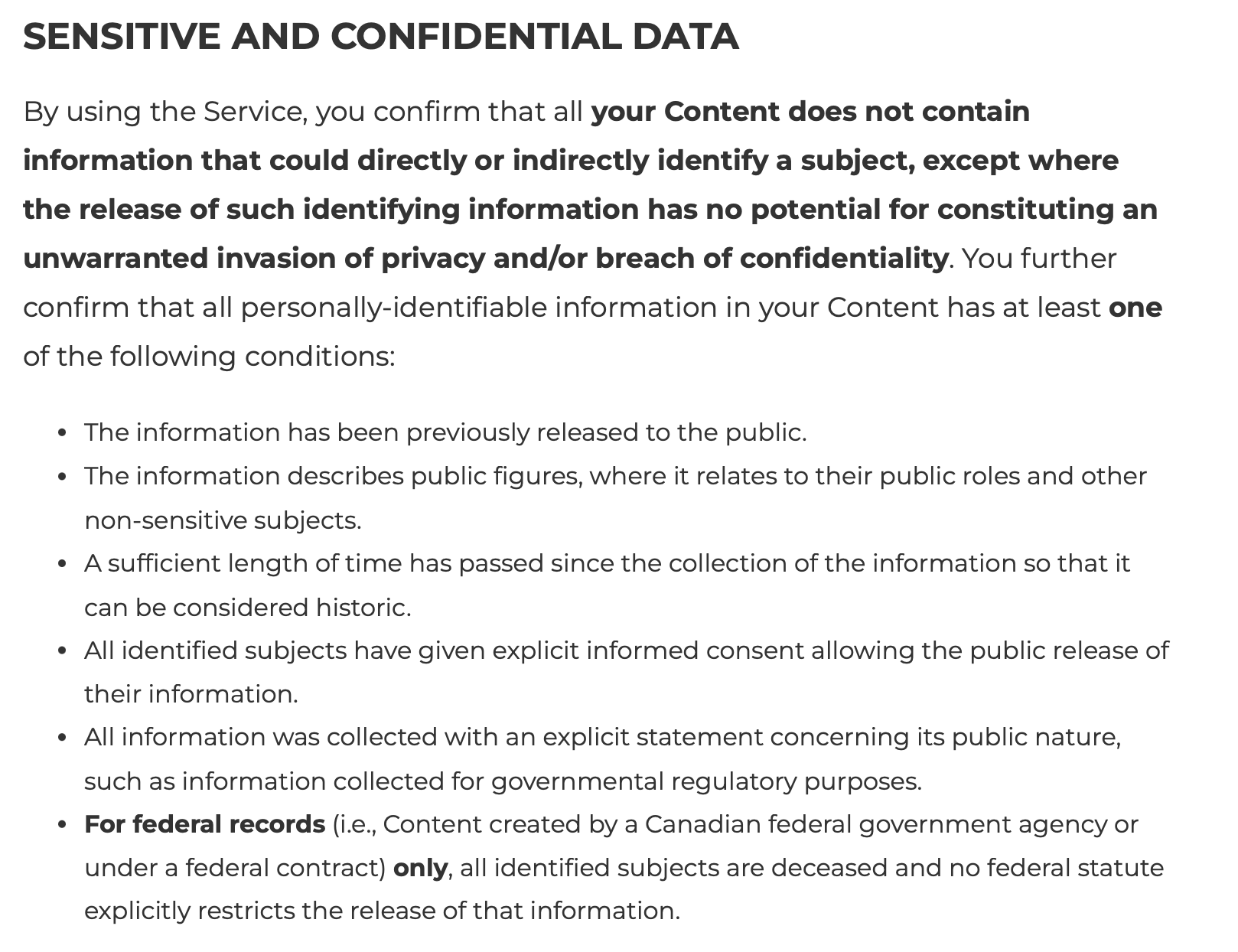
There, the Dataverse is bilingual, well fleshed out, and continually curated (in part by yours truly, me and other staff members). That being said, I don’t speak on behalf of the service – though it nonetheless seems to be me to be a viable approach. Of course discussions would need to be held in order to hash out the details.

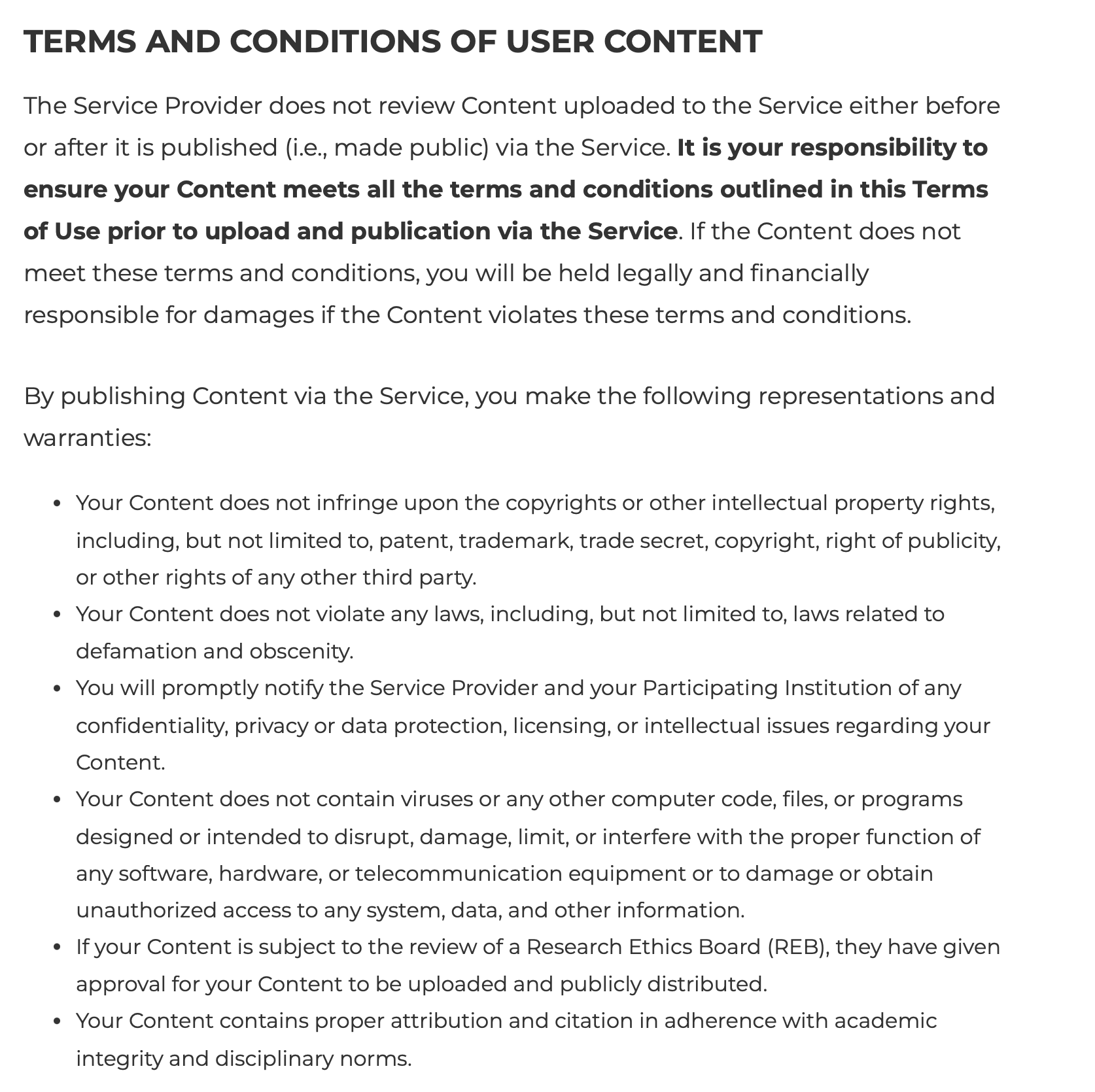
As it stands, Odesi has various Dataverses for a wide range of different data types. Given the nature of the sensitive Metadata and of the present project, not all would be suitable – nevertheless here is the list of existing Odesi Dataverses:

1. Census of Population
2. Statistics Canada Aggregate Data
3. Statistics Canada Public Use Microdata Files (PUMFs)
4. Canadian Public Opinion Polls
5. Statistics Canada Data Liberation Initiative (DLI) Licensed Data
6. **Other Canadian and Government Data**
7. International Data

As readers of the present document may have astutely noticed, (6) Other Canadian and Government Data was bolded to denote an emphasis. It is of my belief that this could be a viable repository for the present endeavour. So let us keep this Dataverse in mind and proceed to other points of discussion.

What Can and “Cannot” Go in Borealis Dataverses (Including Odesi)

Borealis does have extensive documentation about what can and cannot go into their data repositories (this is for very obvious legal reasons). I therefore recommend an in-depth read of [Borealis’ terms of use](https://borealisdata.ca/termsofuse/) page. For the sake of brevity, here are screenshots of the Terms of Use directly relevant to the subject at hand (though of course the entire Terms of Use should be carefully reviewed in order to build a convincing argument/case as to why our goals are justified, safe, and necessary for the preservation of metadata.



Our case has to be air-tight with a clear goal. Data uploaded to Borealis cannot be deleted – it can be [deaccessioned](https://learn.scholarsportal.info/all-guides/borealis/deleting-data/#Deaccessioning-a-Published-Dataset), but not fully deleted. Meaning that it cannot be removed without approval from the highest authority in the Borealis institutional hierarchy.

All this being taken into consideration, Borealis takes almost everything. As an example, a student (I hope not faculty) at Western has been uploading [Harry Potter self-insert fanvids](https://borealisdata.ca/dataverse/esapuridis) in their own Western Dataverse (which *technically* could be infringing on the fanvids’ original owners’ copyright and intellectual property rights, but I digress). The University of Toronto also has a collection of 240+ [pictures of icicles](https://borealisdata.ca/dataverse/Icicle_Atlas). My point is that anything goes in and is minted a DOI.

Copyright and Intellectual Property

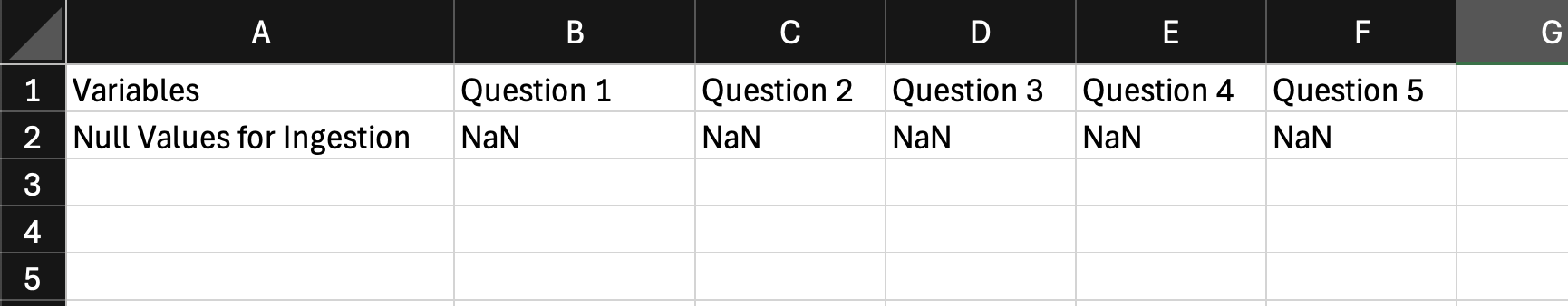
Copyright terms of service are of paramount importance on Borealis (for serious data, at least). To briefly come back to the Harry Potter fanvids mentioned above, they used “CC0 1.0” meaning that it is public domain - which it may or may not be (probably not), but by default datasets in Dataverse are given CC0 1.0 unless specified otherwise. The icicle images, on the other hand, uses “CC-BY 4.0”, which has more constraints on how the data can be used (see [Creative Commons attributions](https://creativecommons.org/licenses/by/4.0/)).

Statistics Canada datasets uploaded to the Odesi Dataverse use a custom template that go beyond Creative Commons rights to address the specific copyrights of Statistics Canada data. I’ll attach it to the appendices (Appendix A for English, Appendix B for French) of the present document. Point is, were we to upload metadata concerning sensitive datasets, we would not only need to strongly uphold Borealis’ Terms of Use, but also ensure that we do not put ourselves in any line of fire for any litigations (or adjacent actions) as this latter would prove problematic for both Borealis and the CRDCN. Provided that these terms are appropriately met, there should be no problem in preserving metadata on Borealis.

Pragmatics of Metadata Findability in Borealis

Assuming that everything aligns with Terms of Use and copyright laws, the upload process is relatively simple. Metadata in Borealis is generated both manually and automatically (assuming the uploaded file is tabular in nature such as .xlsx, .csv, .sav files – but not .RData or .json). This is precisely where we want to make an important decision: do we want a file that is ingested, thus including variable level metadata in the DDI standard XML metadata file, or do we not want that?

It is of my belief that an ingested tabular formatted file would provide the best “findability” and preservation outcomes. To demonstrate this I have conducted a small trial in demo.borealisdata.ca (data here is prone to being deleted and is to test Borealis updates before it ends up in production).

Here is the dataset itself:

It really is as simple as it gets. The first row denotes the variables (imagine that each “Question” actually denotes a Variable of some sort). Note that the “Null” entries are necessary – they can just as well be interchanged for NaN or 0 or “Cheese” or any place holder. However, it is better to just use a string input NaN as it stops the software for calculating summary statistics that it would usually compute for data points in a column. Without these Null values (meaning the dataset is composed of a singular row), the file will not be ingested.

Now here is an example between a non-ingested DDI XML file, and one that is ingested, with this same .xlsx file (note that it would be the exact same outcome were it .csv, or any other – as long as it is not json or RData; this is an issue with Harvard, not Borealis per se).

Non-ingested output:

However if the file is ingested, it also has this following section in addition to the above metadata information (note that it is truncated on the right side and bottom):

This entire entry above is not included if the file is not ingested; this would mean that we would be missing out on a lot of machine readable data were we to not ingest the files. I have already written a python code to exclusively extract variable labels if ever this were needed.

We can also bring this up a notch in terms of metadata quality by including offered response types, but NOT who responded, and how they responded – though I still need to do some testing to see what works and what doesn’t.

Appendices

Appendix A: Terms of Use (English)

## **Statistics Canada Open Licence**

This licence is issued on behalf of Her Majesty the Queen in Right of Canada, as represented by the Minister for Statistics Canada (“Statistics Canada”) to you (an individual or a legal entity that you are authorized to represent).

The following are terms governing your use of the Information. Your use of any Information indicates your understanding and acceptance of the terms below. If you do not agree to these terms, you may not use the Information.

Statistics Canada may modify this licence at any time, and such modifications shall be effective immediately upon posting of the modified licence on the Statistics Canada website. Your use of the Information will be governed by the terms of the licence in force as of the date and time you accessed the Information.

### **Definitions**

“Information” means the compilation of non-confidential results from any Statistics Canada activities, including data files, databases, public use microdata files, publications, tables, graphs, maps, reports and text for which Statistics Canada is the owner or a licensee of all intellectual property rights and made available to you in accordance with this licence, at cost or no cost, either on the Statistics Canada website or by other means as a result of a contract for goods or services.

“Value-added Products” means any products you have produced by adapting or incorporating the Information, in whole or in part, in accordance with this licence.

### **Licence Grant**

Subject to this licence, Statistics Canada grants you a worldwide, royalty-free, non-exclusive licence to:

* use, reproduce, publish, freely distribute, or sell the Information;
* use, reproduce, publish, freely distribute, or sell Value-added Products; and,
* sublicence any or all such rights, under terms consistent with this licence.

In doing any of the above, you shall:

* reproduce the Information accurately;
* not use the Information in a way that suggests that Statistics Canada endorses you or your use of the Information;
* not misrepresent the Information or its source;
* use the Information in a manner that does not breach or infringe any applicable laws;
* not merge or link the Information with any other databases for the purpose of attempting to identify an individual person, business or organization;
* not present the Information in such a manner that gives the appearance that you may have received, or had access to, information held by Statistics Canada about any identifiable individual person, business or organization; and
* not disassemble, decompile or in any way attempt to reverse engineer any software provided as part of the Information.

### **Intellectual Property Rights**

Intellectual property rights, being any and all intellectual property rights recognized by the law, including but not limited to, intellectual property rights protected through legislation, in Value-added Products, shall vest in you, in such person as you shall decide or as determined by law.

Intellectual property rights that Statistics Canada may have in the Information shall remain the property of Statistics Canada. Intellectual property rights that third parties may have in the Information shall remain their property.

### **Acknowledgment of Source**

(a) You shall include and maintain the following notice on all licensed rights of the Information:

Source: Statistics Canada, name of product, reference date. Reproduced and distributed on an “as is” basis with the permission of Statistics Canada.

(b) Where any Information is contained within a Value-added Product, you shall include on such Value-added Product the following notice:

Adapted from Statistics Canada, name of product, reference date. This does not constitute an endorsement by Statistics Canada of this product.

### **Advertising and Publicity**

You shall not include on any reproduction of the Information or any material relating to your Value-added Product, or elsewhere:

(a) the name, crest, logos or other insignia or domain names of Statistics Canada or the official symbols of the Government of Canada, including the Canada wordmark, the Coat of Arms of Canada, and the flag symbol, without written authorization from the Treasury Board Secretariat. Request for authorization from the Treasury Board Secretariat may be addressed to:

[information@fip-pcim.gc.ca](mailto:information@fip-pcim.gc.ca)  
Federal Identity Program  
Treasury Board of Canada Secretariat  
300 Laurier Avenue West  
Ottawa, Canada K1A 0R5

(b) any annotation that may be interpreted as an endorsement by the Statistics Canada of the Value-added Product or that would imply that you have an exclusive distribution arrangement for any or all of the Information or that you have access to any confidential information or information not available to others.

### **No Warranty and no Liability**

The Information is licensed ‘as is’, and Statistics Canada makes no representations or warranties whatsoever with respect to the Information, whether express or implied, in relation to the Information and expressly disclaims any implied warranty of merchantability or fitness for a particular purpose of the Information.

Statistics Canada or any of its Ministers, officials, servants, employees, agents, successors and assigns shall not be liable for any errors or omissions in the Information and shall not, under any circumstances, be liable for any direct, indirect, special, incidental, consequential, or other loss, injury or damage, however caused, that you may suffer at any time by reason of your possession, access to or use of the Information or arising out of the exercise of your rights or the fulfilment of your obligations under this licence.

### **Term**

This licence is effective as of the date and time you access the Information and shall terminate automatically if you breach any of the terms of this licence.

Notwithstanding termination of this licence:

1. you may continue to distribute Value-added Products for the purpose of completing orders made before the termination of this licence provided you comply with the requirements set out in the Acknowledgment of Source clause; and
2. individuals or entities who have received Value-added Products or reproductions of the Information from you pursuant to this licence will not have their licences terminated provided they remain in full compliance with those licences.

### **Survival**

All obligations which expressly or by their nature survive termination of this licence shall continue in full force and effect. For greater clarity, and without limiting the generality of the foregoing, the following provisions survive expiration or termination of this licence: Acknowledgment of Source, and No warranty and no Liability.

### **Applicable Law**

This licence is governed by the laws of the province of Ontario and applicable laws of Canada. Legal proceedings related to this licence may only be brought in the courts of Ontario or the Federal Court of Canada.

**Citation Requirements**: The publishing of analysis and results from research using any of the data products is permitted in research communications such as scholarly papers, journals and the like. The authors of these communications are required to cite Statistics Canada as the source of the data, and to indicate that the results or views expressed are those of the author/authorized user and are not those of Statistics Canada. Permission to include extracts of these data in textbooks must be obtained from the Licencing Section of Statistics Canada's Marketing Division.

**Conditions**: Data Liberation Initiative Community.

Appendix B: Terms of Use (French)

## **Licence ouverte de Statistique Canada**

La présente licence est accordée au nom de Sa Majesté le Roi du chef du Canada, représentée par le ministre responsable de Statistique Canada (« Statistique Canada ») à vous (un particulier ou une personne morale que vous êtes autorisé à représenter).

Vous trouverez ci-après les conditions qui régissent votre utilisation de l'information. Votre utilisation de toute information est une indication que vous comprenez et acceptez ces conditions. Si vous n'acceptez pas ces conditions, il ne vous est pas permis d'utiliser l'information.

Statistique Canada peut modifier cette licence en tout temps, et ces modifications entreront en vigueur dès la publication de la version modifiée de la licence dans le site Web de Statistique Canada. Votre utilisation de l'information sera régie par les conditions de la licence en vigueur à la date et à l'heure où vous avez accédé à l'information.

### **Définitions**

« Produit à valeur ajoutée » : tout produit que vous avez élaboré en adaptant ou en intégrant l'information, en tout ou en partie, conformément aux conditions de la présente licence.

### **Octroi de licence**

Sous réserve des conditions de la présente licence, Statistique Canada vous accorde une licence mondiale, libre de redevances et non exclusive vous permettant :

* d'utiliser, de reproduire, de publier, de diffuser gratuitement ou de vendre l'information ;
* d'utiliser, de reproduire, de publier, de diffuser gratuitement ou de vendre des produits à valeur ajoutée ;
* d'accorder des sous licences conférant une partie ou la totalité de ces droits, conformément aux conditions de cette licence.

Durant l'exécution de toute activité susmentionnée, vous devez :

* reproduire l'information avec exactitude ;
* ne pas utiliser l'information d'une façon qui laisse croire que Statistique Canada vous appuie ou appuie l'utilisation que vous faites de l'information ;
* ne pas présenter de manière inexacte l'information ou sa source ;
* utiliser l'information d'une manière qui ne viole ni n'enfreint toute loi applicable ;
* ne pas fusionner ni lier l'information à toute autre base de données pour tenter d'identifier une personne, une entreprise ou une organisation particulière ;
* ne pas présenter l'information d'une façon donnant l'impression que vous auriez pu avoir reçu ou avoir eu accès à des renseignements détenus par Statistique Canada sur toute personne, entreprise ou organisation identifiable ;
* ne désassembler, ni décompiler, ni effectuer quelque procédé d'ingénierie inverse que ce soit sur les logiciels qui font partie de l'information.

### **Droits de propriété intellectuelle**

Les droits de propriété intellectuelle visant les produits à valeur ajoutée, à savoir tout droit de propriété intellectuelle reconnu par la loi, y compris mais sans s'y limiter les droits de propriété intellectuelle protégés par une législation, vous sont attribués ou sont attribués à la personne que vous désignez ou qui est désignée par application de la loi.

Les droits de propriété intellectuelle visant l'information que possède Statistique Canada demeurent la propriété de Statistique Canada. Les droits de propriété intellectuelle visant l'information qui appartient à des tiers demeurent la propriété de ces derniers.

### **Mention de la source**

(a) Pour tout exercice de vos droits d'utilisation de l'information, vous devez inclure et maintenir la mention suivante :

Source : Statistique Canada, nom du produit, date de référence. Reproduit et diffusé « tel quel » avec la permission de Statistique Canada.

(b) Pour toute information contenue dans un produit à valeur ajoutée, vous devez inclure dans ce produit à valeur ajoutée la mention suivante :

Adapté de Statistique Canada, nom du produit, date de référence. Cela ne constitue pas une approbation de ce produit par Statistique Canada.

### **Promotion et publicité**

Il vous est interdit d'utiliser sur toute reproduction de l'information ou sur tout matériel ayant trait à votre produit à valeur ajoutée, ou ailleurs :

(a) le nom, l'emblème, les logos ou tout insigne ou nom de domaine de Statistique Canada ou les symboles officiels du gouvernement du Canada, y compris le mot symbole « Canada », les armoiries du Canada et le symbole du drapeau, sans l'autorisation écrite du Secrétariat du Conseil du Trésor. La demande d'autorisation au Secrétariat du Conseil du Trésor peut être adressée à :

[information@fip-pcim.gc.ca](mailto:information@fip-pcim.gc.ca)  
Programme de coordination de l'image de marque  
Secrétariat du Conseil du Trésor du Canada  
300, avenue Laurier Ouest  
Ottawa (Canada) K1A 0R5

(b) toute annotation qui pourrait être interprétée comme une approbation du produit à valeur ajoutée par Statistique Canada ou qui sous-entendrait que vous avez conclu une entente de distribution exclusive pour une partie ou pour toute l'information, ou que vous avez accès à des renseignements confidentiels ou non accessibles à d'autres parties.

### **Pas de garantie ni de responsabilité**

L'information est octroyée sous licence « telle quelle », et Statistique Canada ne fait aucune assertion et n'offre aucune garantie d'aucune sorte, explicite ou implicite, relativement à l'information et rejette expressément toute garantie implicite de qualité marchande de l'information ou de son utilité à des fins particulières.

Statistique Canada ni aucun de ses ministres, dirigeants, fonctionnaires, employés, agents, successeurs et ayant droit ne sera tenu responsable d'aucune erreur ni omission dans l'information et ne sera en aucun cas tenu responsable des pertes, blessures ou dommages directs, indirects, spéciaux, conséquents ou autre, quelle qu'en soit la cause, que vous pourriez subir à n'importe quel moment en raison de votre possession de l'information, de votre accès à cette information ou de son utilisation, ou résultant de l'exercice de vos droits ou du respect de vos obligations aux termes de la présente licence.

### **Terme**

La présente licence entre en vigueur à la date et à l'heure où vous accédez à l'information et est résiliée automatiquement si vous enfreignez l'une des conditions de cette licence.

Nonobstant la résiliation de cette licence :

1. vous pouvez continuer de distribuer les produits à valeur ajoutée aux fins de remplir les commandes faites avant la résiliation de la licence, à condition que vous respectiez les exigences énoncées dans la clause de mention de la source ;
2. les licences des particuliers ou des personnes morales auxquels vous avez fourni des produits à valeur ajoutée ou des reproductions de l'information en vertu de la présente licence ne seront pas résiliées à condition qu'ils continuent à se conformer entièrement aux conditions de ces licences.

### **Survie**

Les obligations qui survivent à la résiliation de la présente licence, expressément ou en raison de leur nature, demeureront en vigueur. Pour plus de clarté, et sans limiter la généralité de ce qui précède, les dispositions qui suivent survivent à l'expiration ou à la résiliation de la présente licence : « Mention de la source » et « Aucune garantie ni responsabilité ».

### **Lois applicables**

La présente licence est régie par les lois de la province de l'Ontario et les lois applicables du Canada. Toute procédure judiciaire se rapportant à cette licence ne pourra être portée que devant les tribunaux de l'Ontario ou la Cour fédérale du Canada.

Exigences de citation : La publication d'analyses et de résultats de recherche utilisant l'un des produits de données est autorisée dans les communications de recherche telles que les articles scientifiques, les revues, etc. Les auteurs de ces communications sont tenus de citer Statistique Canada comme source des données et d'indiquer que les résultats ou opinions exprimés sont ceux de l'auteur/utilisateur autorisé et non ceux de Statistique Canada. L'autorisation d'inclure des extraits de ces données dans des manuels doit être obtenue auprès de la Section des licences de la Division du marketing de Statistique Canada.

Conditions: Communauté d’initiative de démocratisation des données