Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

Statistics Canada and Open Data

Bill Joyce[1]

Abstract

This paper is intended to give a brief overview of Statistics Canada's involvement with open data. It will first discuss how the principles of open data are being adopted in the agency's ongoing dissemination practices. It will then discuss the agency's involvement with the whole of government open data initiative. This involvement is twofold: Statistics Canada is the major data contributor to the Government of Canada Open Data portal, but also plays an important behind the scenes role as the service provider responsible for developing and maintaining the Open Data portal (which is now part of the wider Open Government portal.)

Key Words: Open data; Statistics Canada, Government of Canada.

1. Introduction

This paper is intended to give a brief overview of Statistics Canada's involvement with open data. It will first discuss how the principles of open data are being adopted in the agency's ongoing dissemination practices. It will then discuss the agency's involvement with the whole of government open data initiative. This involvement is twofold: Statistics Canada is the major data contributor to the Government of Canada Open Data portal, but also plays an important behind the scenes role as the service provider responsible for developing and maintaining the Open Data portal (which is now part of the wider Open Government portal.)

2. Adopting the principles of Open Data in our normal ongoing dissemination activities

At Statistics Canada over the last number of years, there has been a gradual and continual evolution from printed products to online products and from for-fee products towards free products. In 2012, the agency made a major step in this direction when all online standard data products became free of charge. At the same time, we removed all licensing and royalty fees and with the adoption of an open license, we removed previously existing barriers to the redistribution of our data.

Much of the data found on our website (data found in CANSIM, for example) has traditionally been downloadable and machine readable but we are taking further steps in this direction in the coming months when many of our old style 'data publications' (in HTML and PDF) will be replaced with data base driven tables which will be downloadable. At the same time we will be offering a free API for all users to directly harvest data from our public database of aggregate data at their convenience, 24 hours per day, 7 days per week.

We are able to say that we are respecting the principles of open data in our normal ongoing data dissemination activities.

The definition of open data that we are using is as follows: open data are free of cost, free of barriers to redistribution and machine readable. This is consistent with how open data is defined around the world, but some organizations go

[1]Bill Joyce, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, K1A 0T6 (bill.joyce@canada.ca)

further and include the concept of "linked data". In Sir Tim Berners Lee's five star rating for open data, the 4th and 5th stars relate to linked data. Data formatted for the semantic web is another label for this concept and formats such as RDF and SPARQL are used.

We are not working towards the production of data which are formatted for the semantic web. Data users are not asking for data in these formats and some would argue that the concept is somewhat academic and theoretical. On the other hand, we need to ensure that we will not realize at some future point that the demand for data to feed the semantic web is real and that we have "missed the boat". The position of the agency is that we will watch the development of linked data, but that we will not invest in this format at the current time.

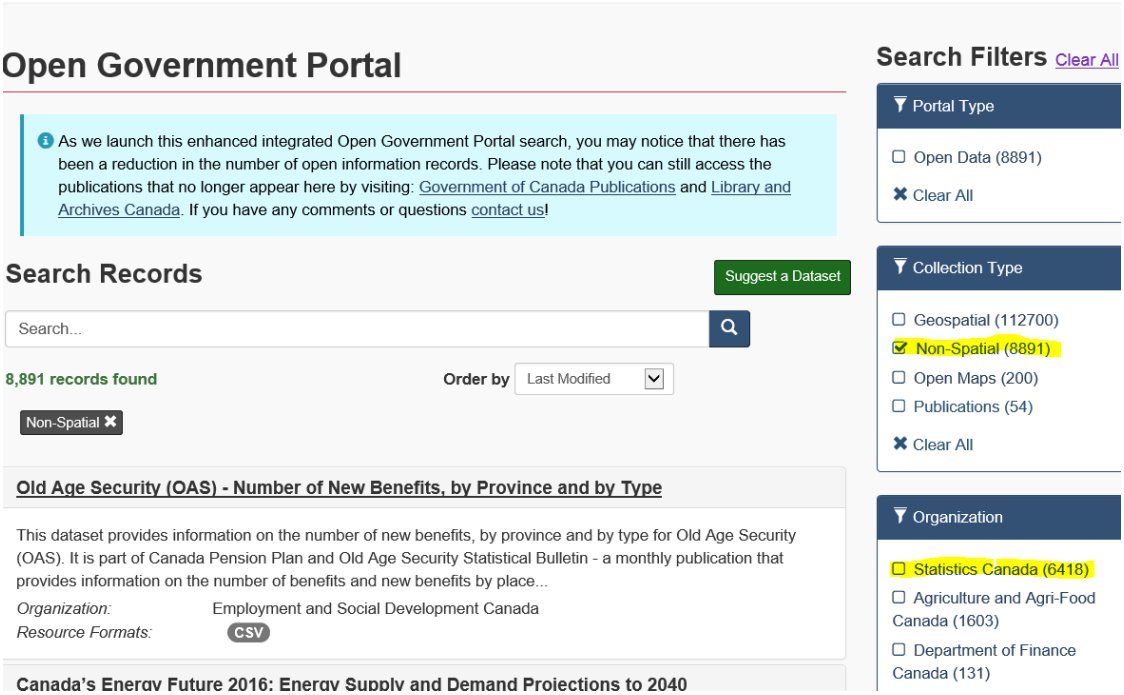
Our microdata files, even our Public Use Microdata Files, are not directly available online, nor are they technically considered as open data. This is not to say that we do not want to increase the accessibility of microdata. We have a range of available options to ensure that microdata are made available to researchers who need it through the use of remote data access, secure research data centres, etc. Our Public Use Microdata Files, however, come with strict licensing restrictions which include, for example, a prohibition on linking or merging files. For this reason a true open license does not apply and we do not consider this data to be a component of our open data inventory.

Custom data tabulations are provided to our users on a cost recovery basis. This output is still provided with an open license, and no licensing or royalty fees are applied – but the cost is based on the time it takes our staff to produce the custom output. We still feel that we can say that we are respecting the principles of open data as we always seek to provide as much standard free data as possible.

3. Statistics Canada as a main data provider to the government wide open data portal

Since the first pilot of a federal government-wide Open Data portal in Canada, Statistics Canada has been the main data provider.

The following screen shot from open.canada.ca shows that 72% of the non geo-spatial data files are supplied by Statistics Canada.



As a statistical agency, our data are always vetted for confidentiality and are designed to be published. As such we have a great advantage over other federal departments who are, for the first time, now being asked to adopt an 'open by default' approach to their data files.

The federal Open Data portal (now part of the wider Open Government portal) does not hold a duplicate copy of any of our datasets. Instead, the portal holds a central catalogue or registry and each record includes a pointer back to our servers where the original downloadable data files are housed.

Regarding formats, most of Statistics Canada's data are available in both CSV and XML (SDMX-ML).

4. Statistics Canada as a service provider for open.canada.ca

The Treasury Board of Canada Secretariat is the federal government department responsible for the Open Government portal (which originally started as an open data portal). In their search of a lead for the technical implementation of the portal, it was deemed that Statistics Canada had the appropriate experience and expertise required.

Consequently, in addition to being a data provider, Statistics Canada also provides the computer science professionals for the development and maintenance of the federal government portal and coordinates the provision of the IT infrastructure required to support the portal.

A quick tour of open.canada.ca will provide an idea of the types of services available. From a technical perspective, we use the content management system Drupal which supports the publishing of static content, blogs, commenting and rating, etc. The main data catalogue uses a technology called CKAN.

5. Conclusion

It is clear that the concept of open data is a natural fit with the mission and mandate of national statistical agencies. Statistics Canada is committed to ensuring that Canadians have barrier-free access to key information that they require to function effectively as citizens and decision makers in a rapidly evolving world.