**Evaluating Canada's Federal Open Data Portal for Criteria of Effective Use**

Written Assignment 3: Data Analysis Exercise

Submitted by: Cody Skinner
December 16, 2013
Submitted to: Helen Hambly Odame
CDE 6410: Readings - Capacity Building

**Summary**

This paper begins with an introduction to open data and the problem of measuring its impact. After a review of the priorities in approaching impact measurement put forth by leading open data researchers and practitioners, the paper looks to the same community to identify the optimum pre-conditions for successful open data use. A selection of the criteria for effective use that can be addressed on the supply side of open data are then operationalized, with indicators selected for each from available data from the Canadian federal open data portal. Finally, download counts for a sample of datasets from the federal portal are tested against these indicators to discover if datasets that meet the supply-side criteria for effective use are also in high demand.

**Background**

The most popular definition of open data comes from the Open Knowledge Foundation's 'Open Definition' (OKF, 2012). To meet the criteria of this definition, data must be available as a whole and at no more than a reasonable reproduction cost; it must also be available in a convenient, machine-readable and modifiable format; its release must permit re-use and redistribution, including mash ups with other datasets; it must be available for use by all, commercial or non-commercial, and regardless of political or social interest (OKF, 2012).

The motives for governments to open their data include benefits such as increased transparency, accountability and better measurement of the impact of policies; more effective and efficient government; participation and self-empowerment; and economic, academic and social innovation (Davies, 2013a; OKF, 2012; Hujboom & Van den Broek, 2011; Davies, 2010). The 2013 Open Data Barometer report uses three components to gauge a country's readiness for using open data to these ends: government capacity and the presence of government commitments to open data; citizen and civil society freedoms and engagement with the open data agenda; and, resources available to entrepreneurs and businesses to support economic re-use of open data (Davies, 2013a). In another framework, Alonso

(2013) describes six dimensions of open data readiness, include legal, political, social, economic, organizational and technical capacity.

Government data is valued for reliability, standardization, consistency and comprehensiveness (Lakomaa & Kallberg, 2013; Davies, 2013a). Innovators are much more likely to invest time and resources in projects using data when they are assured about its continued availability, while social applications of data are often reliant on frequent updates and timely access (Davies, 2013a). Even so, open data disrupts the government's traditional role as the data owner (Helbig, Cresswell, Burke and Luna-Reyes, 2012; Davies, 2010). Helbig et al. (2012), suggest that all stakeholders be regarded as stewards of open data with a joint responsibility to assure the accuracy, validity, security, management, and preservation of data holdings. In the authors' words, "the governance of open data initiatives involves creating policies, business processes, social processes, technologies, standards, meaning and interpretation, and adding value" (Helbig et al., 2012, p. 14).

The wide ranging impacts of open data make their measurement a difficult proposition (Davies, 2013b). Although the costs of producing and publishing data can be measured, there is as of yet no widely accepted model for evaluating the benefits of open data (Shakespeare, 2013; Hujboom & Van den Broek, 2011; Davies, 2013b). Although various methods have been suggested to generate indicators of the impacts of open data, "none of the proposed methods are comprehensive or rigorous enough to encompass all the key aspects, and potential impacts, of assessing an open data initiative - nor to provide effective coverage of developed and developing nations" (Davies, Perini and Alonso, 2013, p. 27).

Huijboom and Van den Broek (2011) argue that the precise economic impact of open data remains largely unclear, while there is even less evidence of the effects of open data on participation, accountability and transparency. While there is more data available to citizens, whether or not that data holds information relevant to the issues the public is interested in is another matter (Halonen, 2013). This is problematic as measures of the impact of open data should be central to informing decisions

around the design and implementation of open data initiatives (Ubaldi, 2013; Jetzek, Avital, & Bjørn-Andersen, 2012; Helbig et al., 2012).

If economic benefits are the easier of open data's acclaimed benefits to measure, any wider social or political effects are more difficult to prove, requiring impact analysis using non-economic or non-quantitative means (Halonen, 2013). Halonen (2013) points out that the overall number of data users may even be irrelevant, with the more important indicators found among those actually making use of open data with relatively high impact (Halonen, 2013). In the author's view, measuring transparency is not simply a question of the actual amount of information available, but also a question of equality in the accessibility and usability of that information. An assessment of open data's effects on participation must assess whether improved access to an increased amount of data will change citizens' behavioural patterns. The author suggests that measuring the effects of open data on participation may even require a re-examination of the concept of participation itself, noting there are "many calls for redefining participation from a certain duty-based citizenship to more of an engaged form of citizenship" (Halonen, 2013, p. 92).

Jetzek et al. (2012) suggest open data intiatives should assess the potential value open data may generate. The authors contend that open data use can be conceptualized as a value network, in which value is co-produced through a process that offers due consideration for the different and divergent interests of all collaborating partners (Jetzek et al., 2012). From the perspective of the private sector, this value can be accounted for in cost-savings, both direct and indirect, or opportunities to generate new revenue. Ubaldi (2013) suggests impact measurements are key to providing practitioners the incentive needed to move from proof-of-concept to production applications of open data to begin with. On the 'social' side of open data, Jetzek et al. (2012) suggest looking for effects on existing indicators of equality or life expectancy. However it may be a long time before significant effects are seen using these measures, nor do they account for new or unexpected benefits from open data use, with some researchers arguing that there "is no linear progression from data, to decision-making, to impact – but

rather that these are in on-going interaction" (Davies, Perini and Alonso, 2013, p. 13).

Davies, Perini and Alonso (2013) argue a socio-technical approach to assessing open data impacts would provide an understanding of the flow of data from open data initiatives to potential users through a range of technical and social intermediaries. It would reveal how global standards, platforms, infrastructure and 'eco-systems' of open data affect local contexts and how the benefits from open data initiatives are distributed. The authors recommend researchers make use of case studies towards these ends but contend that it is also important to move beyond local findings to assess open data initiatives at the macro level (Davies, Perini and Alonso, 2013). By doing so, it will become possible to "understand commonalities between cases of open data publication and use, and to uncover common mechanisms through which open data may be involved in bringing about impacts" (Davies, Perini and Alonso, 2013, p. 12 ).

Open data researchers and practitioners seem much more certain about the necessary preconditions for the success of open data than they are of how to measure it. Gurstein (2011) lays out a number of requirements for what he calls the 'effective use' of open data. These include an accessible and sufficient internet connection, hardware and software; the skills to use those tools; the data in a useful format; sufficient knowledge to make sense of the data; resources sufficient for translating data into activities for local benefit; and the legal, regulatory or policy regime needed to use the data (Gurstein, 2011). Meeting the criteria of this framework, in the author's view, would ensure opportunities and resources for translating open data into useful outcomes are available for the widest possible range of users (Gurstein, 2011).

Halonen (2013) contends that the failure to account for these criteria presents a potential risk of widening the gap in access open data is, in part, meant to close. Despite claims that open data will provide new actors with the means to participate in public debate, it can be argued those most likely to take advantage of open data are already empowered and engaged individuals with a particular technical skill set (Davies, 2010; Halonen, 2013). Furthermore, Halonen (2013) suggests that a public that is

arguably already apathetic to the information currently available to them may be even more put off by the challenge of accessing and analyzing raw data.

Even before the advent of open data, Sawicki and Craig (1996) had begun to recognize major challenges in the use of data for social innovation and change. They identify a lack of technical sophistication among community groups, and the eventual transformation of data into analysis that can affect policy. “[E]ncyclopedic data dumps do not have much impact. They are not issue- or policy-oriented, and often do not serve as even basic references” (Sawicki and Craig, 1996, p. 518). The authors view the adoption of data-enabled strategies within community organizations as a further challenge: “The degree to which the products of information technology are produced for community organizations as opposed to being produced by community organizations themselves is an empowerment issue” (Sawicki and Craig, 1996, p. 518).

While intermediaries are able to process data into platforms and products with social and economic value for others (Davies, 2013a), some advocates are calling on open data curators to make open data inclusive beyond app developers and data scientists (Meijer et al., 2012; Davies, 2010; Desouza & Bhagwatar, 2012). Davies (2010) contends that much of open data use, and some of its greatest impact, is not the result of sophisticated software development or visualization, but of being able to directly identify facts within datasets and process raw data into information.

For some this translates to a need for more accessible tools to easily manipulate data into summaries and visualizations in a customized way that meet a user's needs (Halonen, 2013). Others reach deeper, calling for the release of government open data to be accompanied by complementary programming to increase data literacy (Davies, 2012; Shakespeare, 2013). While “technology platforms, common standards and open licences all play a key role in making data easier to access and simpler to process, they operate against a wider backdrop of organization and social arrangements, power dynamics, and market conditions which may or may not be favorable to the use of data by different agents” (Davies, 2012, p. 1).

While Gurstein's (2011) own work refers mostly to what he calls the 'demand' side, he admits that aspects of effective use are dependent on the supply side of open data. The elements of the Gurstein refers to are often matters of how the data is preprocessed and presented. This includes formatting, tagging and metadata, but is also a matter of how presentation can affect its interpretation, for example by choice of interval in time series data, or the choice to include certain measurements over others.

The impacts of these characteristics will vary between users and how they intend to use the data, with different requirements needed of the data depending on the desired outcome (Davies, 2013a; Gurstein, 2011). Helbig et al. (2012) call on open data curators to engage with citizens, developers and other branches of government to seek a greater understanding of their demands for data. Each stakeholder will have their own distinct requirements regarding data quality, file formats, metadata and regularity of updates. The demands of one set of users may conflict with another and, without proper consideration, lead to the pursuit of competing goals within the overall strategies of data agencies (Helbig et al., 2012).

The quantity of data made available by governments does not amount to much if that data is not actually in demand by users (Davies, 2013a). The 2013 Open Data Barometer global report laments that in most countries key data sets for entrepreneurship and policy analysis are not available as open data, and when published are in non-standard formats (Davies, 2013a). Selecting data sets for release is more realistically a matter of balancing resources, time and effort. Pre-existing good data management practices at agencies supplying data will reduce the cost and effort needed to make data available and increasing the probability it will be easy to reuse (Helbig et al., 2012). As Davies (2012) points out, the datasets governments make available are “the product of past political choices, resource constraints and practical considerations” (Davies, 2012, p. 3).

However, some open data advocates fear that if one waited for data to be made perfect before being posted for public use, it would never happen. In a call for governments to place data online in

any format, Tim Berners-Lee has set out a model for open data publication called the ‘Five Stars of Linked Data’ (Openness Rating, 2013). Each level on Berners-Lee's scale represents a step toward making it easier to use data with other data. Interoperability is a key factor to innovation through the use of open data, and denoges the ability to combine different datasets in 'mash-ups' to reveal new or more comprehensive results. In addition to standard licences that allow data from different sources to be used together, interoperability is also affected by file format, metadata, and common data definitions (Helbig et al., 2012; OKF, 2012; Peristeras et al., 2009; Uhlir & Schroder, 2007). The idea behind the model is for governments to post data online even if it is imperfect, and improve it as time goes on (Davies, 2013a). The lowest level simply requires the data to be covered by an open licence, the next level that the data be machine-readable, and the third that the machine-readable format be non-proprietary. The four and five star levels of the model incorporate the use of linked data, which allows users to connect disparate data across the web using unique identifiers (Openness Rating, 2013).

Helbig et al. (2012) would also argue that sufficient context for the supplied data must be provided for an open data set to be of any use, stating “data does not exist in the wild; it is deliberately created by socio-technical processes” (Helbig et al., 2012, p. 13). Context, as per the authors' formulation, is information related to the environment from which data is acquired or extracted, encoded, and typically used to impact government and public life. Failure to include context, the authors argue, can lead to conflicts in overall meaning, misunderstanding of data elements and a lack of use due to uncertainty over the value of the information a dataset holds (Helbig et al., 2012). Different users and audiences will require different information, and so future-proofing data by contextualizing it for a diverse set of interests is important. By ensuring data has sufficient context for use by various audiences and users, open data initiatives can contribute to overall public value creation (Helbig et al., 2012).

**Canada's Open Data Portal**

In June, 2013, the Canadian federal government signed on to the G8 Charter on open data. The main principles of the charter include: the majority of government data should open data by default; a commitment to high quality and quantity of data; a commitment to releasing data that is usable by way of licensing and formatting; a commitment to releasing data that can used for improved governance; and a commitment to the use of data for innovation by increasing data literacy (FATDC, 2013). To complement the announcement, the federal government simultaneously made a number of modifications to their open data program. First, an updated online federal open data portal was launched. This offered better search capabilities, the ability to rate and comment on datasets and technical information on the use of the site's API.

In addition to the portal, and perhaps of greater importance, the government released a new Canadian Open Government Licence which was applied to all data sets on the new federal portal (“Federal open data portal revamp”, 2013). Licences are important for interoperabilityin a legal sense, determining how data can be used. As Eaves (2011) points out, there is no point in using open data if you are unable to share your findings with others. The new licence was meant to address criticism of the existing licence and increase interoperability of federal open data assets (Scassa, 2013; “Open Government Licence Consultation Report”, 2013). The original licence was based on that of the UK, and the new version has removed language that does not apply under Canadian law (Scassa, 2013). The licence's authors chose to leave in requirements for attribution of the data source, although some commentators had felt it should be removed (“Open Government Licence Consultation Report”, 2013). While a few minor differences remain, the licence has now more or less been adopted by the provinces of British Columbia, Alberta and Ontario and a number of municipalities including Vancouver, Edmonton and Toronto (City of Vancouver, 2013).

While the federal government appears to be taking major steps toward opening its data, and as Canada ranked 8th in the world in the 2013 Open Data Barometer report (Davies, 2013a), the federal

initiative still has critics among open data experts within the country. The federal data portal is reportedly lacking in data sets that relate directly to government transparency and accountability, while information is said to be lacking the organization needed for national analysis and comparisons (Davison, 2013). The government's celebrations of its open data initiative have also been overlapped by accusations of muzzling scientists from reporting research findings (Davison, 2013). These accusations leave serious questions concerning how selective the federal government has been in opening its data.

**Sampling, Operationalization of Concepts and Methodology**

The review of literature above has two major focuses, measuring the impacts of open data, and the necessary preconditions for the effective use of open data. One of the major takeaways from this discussion is that simple quantitative measures will not capture the impacts of open data. Although they are not indicators of impact, this paper intends to test some of the simplest measures of open data use available on the Canadian federal open data portal, download counts. It is easy to cite these figures as measures of open data's success, but they are indicators of demand for specific data and should not be confused as measures of the actual success or impact of open data.

While it is beyond the resources of this paper to assess how datasets are being used after they are downloaded, it is possible to evaluate the extent to which datasets in high demand meet the supply-side criteria set out above for effective use. This is, of course, only half of the effective use equation, but, unlike download counts, the extent to which these criteria are met could arguably be taken as an indicator of the potential of a dataset to be used with actual impact. As such, it seems worth asking if the datasets that come closest to meeting those criteria are in high demand.

At the time of the federal open data portal's relaunch, then Minister of Citizenship, Immigration and Multiculturalism, Jason Kenney proudly cited immigration data as making up the top six data sets downloaded from the site (“Canada launches”, 2013). In November 2013, datasets from Citizenship and Immigration Canada continued to be popular, making up up nine of the top ten downloads on the

federal open data portal. As a sample, this paper will use 37 datasets (n=37) from the federal open data portal supplied by Citizenship and Immigration Canada (CIC). In addition to the popularity of these datasets, this selection was made in part due to the process for obtaining download counts. While the federal open data portal lists downloads counts for the top ten downloaded datasets from the previous month, these counts must be requested for all other datasets on the portal through email from the Chief Information Officer Branch within the Treasury Board of Canada Secretariat. To ensure a timely response from the responsible agency, simplified parameters for the request were sought. The main question of this study asks if data with high potential for effective use is in high demand. Because CIC datasets were the most demanded on the portal, it was decided that information for these datasets be requested.

This convenience sampling, however, is not without trade-offs. This cross-sectional study must first and foremost be treated as a case study with limited external validity. While datasets within the sample vary in their demand, it should be assumed that the large proportion of highly demanded datasets in the sample make it unrepresentative of the larger population. Furthermore, because the data is all supplied by the same agency, a second concession is a homogeneity in observations for certain variables in the sample related to criteria for effective use. These will be addressed as those criteria are operationalized below.

As previously stated, download counts will be used as direct indicators of demand for data. The information returned by the Chief Information Officer Branch included counts for total and unique user downloads. In an effort to be more stringent, this research will only use the unique download count as a measure of user demand. Concepts to the supply-side criteria for effective use have been operationalized using metadata and the contents of the actual datasets. A brief description of how this data was obtained is in order. The metadata available on the open data portal for each dataset within the sample has been recorded into a table. This was done by the researcher, and thus opens risk for human error. Upon examining the datasets in closer detail, it became evident that some of the metadata

supplied on the portal pages was erroneous. Corrections were made to data for analysis, but the presence of this inaccurate information has also been noted for further analysis.

From the literature related to supply-side criteria for effective use of open data, this study will focus on operationalizing the concept of context described by Helbig et al. (2012) and by Gurstein (2011). Admittedly, the study would have sought to examine the attributes identified by Tim Berners-Lee's Five Stars of Linked Data, however it is a quirk of the sample that all datasets are in the same, proprietary format, and thus earn an identical two star rating. Any future analysis should use an expanded sample varied in this respect and more representative of the overall collection of data available on the portal.

As per the description of context offered by Helbig et al. (2012), metadata and the content of each dataset will be examined for the inclusion of an explanation of data acquisition or extraction; explanations of complex variables and terms used in the data; and, information related to the data's typical use. These variables are discrete and nominal, coded for the inclusion or omission of each category of contextual information. Admittedly, assigning values for these indicators is a subjective process, despite the guidelines set out below. An in depth qualitative analysis of the relative strength or weakness of these items is outside of the scope of this study, but should be pursued in future work.

Explanation of the data collection process will be regarded as included if it is made clear in general terms how the data is derived from the overall function of the agency. Some datasets in the sample are more granular than others (eg. breaking down various application figures by processing office). This will not be taken into account as it is not so much a failure to explain data collection as it is related to the actual substance of the dataset.

Taking the goal of 'effective use' to mean use by the widest audience possible (Gurstein, 2011), this assessment will consider data variables or terms that need defining to be those that are not in common use, not obvious to someone unfamiliar with Canadian citizenship and immigration issues, or with a definition specific to the operations of CIC. To account for some level of nuance, this variable

will use the three following possible categories: first, specialized terms are included in the data with no explanation provided; second, no specialized terms are used; and, third, explanations for specialized terms have been provided.

Information related to the data's typical use will be assessed as any information related to how the data is used in decision-making from day to day operations to high level policy, or even examples of the data's use by citizens, civil society or the non-profit sector.

The operational definition of context will be extended to incorporate issues identified by Gurstein (2011) including formatting, tagging, metadata, update intervals and the overall period of coverage. As noted above, there are inconsistencies between the metadata and the content of the datasets. As such metadata will be examined for accuracy in three areas: frequency of update, chronological range, and accuracy of description of content. It will be treated as a discrete, nominal variable coded as consistent or inconsistent with the actual contents of the dataset.

All datasets on the federal portal are tagged to aid the search for data of interest, however, some tags are better than others. This paper will not attempt to examine which terms are more likely to be searched than others, this is a topic onto itself and is outside of the scope of this paper. On a more mundane note, however, tags for many of the sampled datasets are dysfunctional due to improper formatting. As such, tags will be examined for formatting mistakes and treated as a discrete, nominal variable coded as containing mistakes or not.

Finally, the total time period covered by each dataset will be examined, as will the intervals at which updates are provided. Unlike all other variables in the operationalized definition of context examined in this paper, these variables could be treated as interval or ordinal data. However, all but three datasets initiate coverage either from the year 2008 until the present, or the year 2012 until the present, and all but three datasets are updated on an annual or quarter-annual basis. As such, while it is technically possible to treat these as higher level data, they will be categorized and treated as nominal data for the sake of simplifying analysis and comparing results to those found when testing other

variables for predictive relationships with download counts.

While nominal data can only be tested for predictive relationships, it may still be worth considering the four criteria for establishing the validity of a causal relationship (time order, nonspuriousness, co-variation, and theory). First and foremost, many of the variables being examined here are truly nominal level data and thus cannot co-vary. While time order can be established for those variables the user is aware of through metadata or is affected by in their search for data, nonspuriousness is next to impossible to establish. Overall, users most likely arrive at the data portal already knowing the type of information they are seeking. Information providing context for the data would more likely reconfirm the user's demand, with true causation intiated by an unknown preceding variable (Berry & Sanders, 2000). Any workable theory of causation would have to assume a user arrives at the portal not knowing what information they are seeking. In any case, causation is not necessarily relevant to the question posed in this paper, which is whether the datasets that come closest to meeting the criteria for effective use are in high demand.

After a brief run through descriptive statistics for the sample, contingency tables will be used to test for any association between the variables identified above and download rates. In addition to summaries of the percentage differences, chi-square tests will be used to evaluate the level of statistical significance of any association evident in the tables. Finally, should those relationships prove statistically significant, Cramer's V will be used as a measure of the level of predictive association between the two variables.

**Analysis and Key Findings**

As evidenced in the histogram and box plot below (Figures 1 and 2, respectively), the distribution of download counts is positively skewed and quite obviously not a normal distribution. The mean of the distribution is 79.41 downloads with a standard deviation of 24.96 and a range of 779, but these figures are misleading given the presence of a few cases with exceptionally high download rates. The median

and interquartile range are more resistant to these outliers and can thus provide a better picture of the distribution in these cases (Hartwig, 1979). This particular distribution has a median of 19 and an interquartile range of 74. Despite it being much smaller than the mathematical average of the distribution, as the 50$^{th}$ percentile the median score will be used to divide the distribution of download counts into categories of 'low' and 'high' for use in contingency tables.

**Figure 1. Histogram of Download Counts**

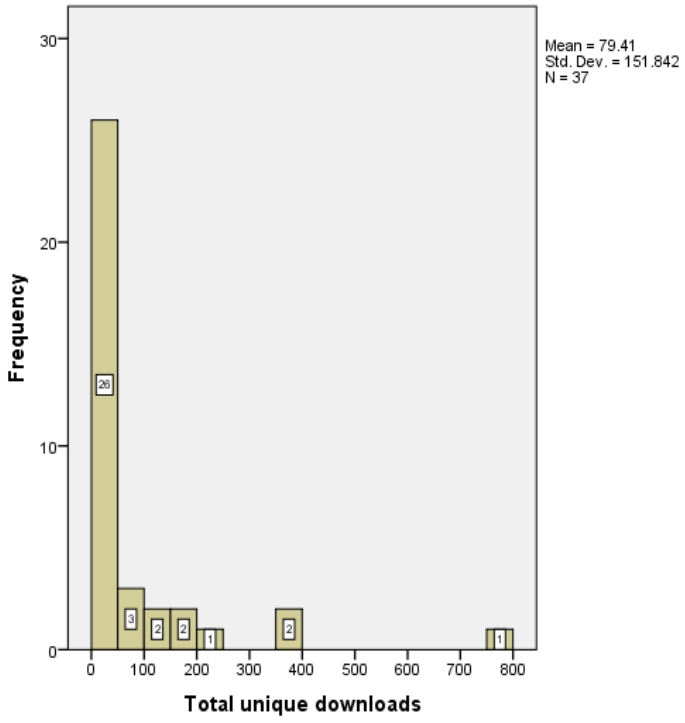
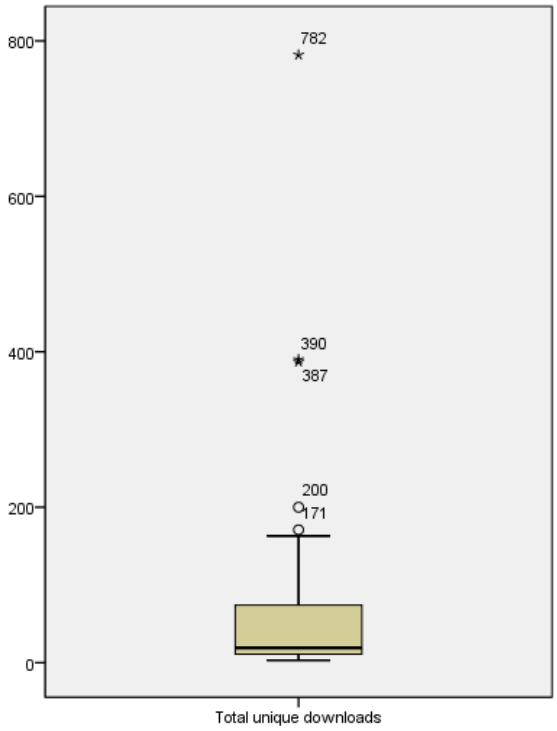



**Figure 2. Box plot of Download Counts**




Upon collecting and examining the data, it became apparent that the values for some variables were uniform. This is unsurprising given that the same agency supplied all of the datasets in the sample. Zero datasets had metadata explaining how the datasets were typically used by the agency or by external agents, while 100% of the datasets had metadata explaining how the data was acquired. As such, these two variables were discarded from analysis. The final variable related to context saw more variation: just 10.8% of all cases (n=37) provided explanation of terms in the dataset, though only 37.8% of all cases were deemed to contain complex terms. Meanwhile, 60% of datasets in the sample

had improperly formatted tags. Frequency tables for all variables can be found in Appendix. Hypotheses for association between the variables and level of demand, and their tests results are discussed below.

**H1: Datasets that include definitions for complex terms, or do not contain complex terms, are more likely to be in high demand.**

While it's likely that knowing a dataset has a clear definition of terms would make one more inclined to download it, it is also likely that if data is in high demand, users would also ask for the context needed to use the data properly. Finally, given that easily comprehended data should be accessible to a wider audience, one would also expect higher demand for this category.

However, while cross-tabulation did reveal a statistically significant association and a moderate predictive relationship using Cramer's V, the nature of the predictive relationship was completely as hypothesized. While the actual number of observed cases is very low, datasets that included explanations for complex terms all fell within the category of high demand. At the same time, 80% of datasets that left complex terms undefined also fell in this category. High demand for complex data could be explained by a users with specialized knowledge and already familiar with CIC definitions, but it is impossible to tell from the available data. With respect to the question of whether data in high demand meets the supply-side criteria for effective use, descriptive statistics have already shown that data definitions are largely absent from all datasets in the sample.

**H2: Datasets with accurate metadata are more likely to be in high demand.**

Given that the metadata of datasets with higher demand are viewed more often than those with lower demand, one would expect inaccuracies to be reported and corrected. However, cross-tabulation reveals an almost identical proportion of inaccurate metadata for datasets in low and high demand, while a chi-square revealed no significant association between accurate metadata and demand.

**H3: Datasets that are formatted correctly are more likely to be in high demand.**

Again, if a dataset is in high demand, there should be more chance that tag formatting mistakes would be reported and corrected. Conversely, proper tagging makes data easier to find and thus to download. Chi-square reveals a statistically significant association (at the .013 level) between tags without formatting mistakes and demand for datasets, while Cramer's V suggests a moderate to low level predictive relationship. Judging by the percentage difference, correct tag formatting makes a difference of 42% in the ability to predict data demand within this sample.

**H4: Datasets that cover a greater amount of time are more likely to be in high demand.**

One would suspect that data that goes back a greater number of years would be of greater value. However, the cross-tabulation reveals little difference between the levels of demand for data covering the past 18 months and data covering the last 66 months. Chi-square does not reveal a statistically significant relationship and the null-hypothesis cannot be rejected.

**H5: Datasets that have shorter intervals between updates are more likely to be in high demand.**

Data that is regularly updated is much richer, and thus one may think it would be in higher demand. Again, however, the cross-tabulation fails to reveal any major difference between demand for data updated quarterly and data updated annually. Without a statistically significant relationship, the null-hypothesis holds. The failure of both the period of coverage and interval of update to have any significant association with demand for the data leads one to suspect the content of the data is the most important quality in determining demand.

**Conclusion**

This study has faced major limitations in the availability of relevant data and by its choice of sample.

However, while its findings are limited, it does provide some insight for future research. First, there is little evidence to suggest that where the supply-side criteria exist for effective use, demand will be any higher than it would typically. However, it is quite evident that these criteria are largely unmet in Citizenship and Immigration Canada data. As such the theory should be further tested using a larger sample that is representative of the larger federal open data catalogue. Furthermore, the failure to find an association between data demand and these criteria says nothing about their validity as pre-requisites for effective use. If one thing is evident from these findings, it is that demand for data largely comes down to the subject and substance of the information. To truly test supply-side criteria for the effective use of open data, one will still have to look towards the data's application and impact.

**References**


Alonso, J. M., Boyera, S., Grewal, A., Iglesias, C., & Pawelke, A. (2013). *Open Government Data Readiness Assessment Indonesia* (pp. 1–47).

Openness Rating. (2013, June 5). *data.gc.ca*. Retrieved from: http://data.gc.ca/eng/openness-rating

Berry, W.D., & Sanders, M.S. (2001). *Understanding Multivariate Research: A Primer for Beginning Social Scientists*. Boulder, CO: Westview Press.

Canada Launches Next Generation Open Data Portal . (2013, June 18). *Citizenship and Immigration Canada*. Retrieved from: http://www.cic.gc.ca/english/department/media/releases/2013/2013-06-18.asp

City encourages wider use of open data, adopts Open Government Licence. (2013, September 24). *City of Vancouver.* Retrieved from: http://vancouver.ca/news-calendar/city-adopts-open-government-licence.aspx

Davies, T. (2010). *Open data, democracy and public sector reform: A look at open government data use from data. gov.uk*. Practical Participation.

Davies, T. (2012). Supporting open data use through active engagement. *W3C Using Open Data Workshop, June 2012*. Brussels.

Davies, T. (2013) *Open Data Barometer: 2013 Global Report*. London: Open Data Institute and World Wide Web Foundation.

Davies, T. (2013, November 14). Open Data and Improving Governance: issues of measurement. Retrieved from: http://www.opendataimpacts.net/2013/11/open-data-and-improving-governance-issues-of-measurement/

Davies, T., Perini, F., & Alonso, J.M. (2013). *Researching the emerging impacts of open data: ODDC conceptual framework*. London: World Wide Web Foundation and ODDC.

Davison, J. June 19 2013.How open is Ottawa's new 'open data' website? *CBC News*. Retrieved from: http://www.cbc.ca/news/technology/how-open-is-ottawa-s-new-open-data-website-1.1373680

Desouza, K. C., & Bhagwatwar, A. (2012). Citizen Apps to Solve Complex Urban Problems. *Journal of Urban Technology, 19*(3), 107-136.

G8 Open Data Charter. (2013). *Foreign Affairs, Trade and Development Canada (FATDC).* Retrieved from: http://www.international.gc.ca/g8/open_data-donnees_ouvertes.aspx?lang=eng

Federal open data portal revamp aims to encourage apps. (2013 , June 18). *CBC News*. Retrieved from: http://www.cbc.ca/news/technology/federal-open-data-portal-revamp-aims-to-encourage-apps-1.1310665

Gurstein, M. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday. 16*(2).

Halonen, A. (2012). *Being Open About Data: Analysis of the UK open data policies and applicability of open data*. London: The Finnish Institute in London.

Hartwig, F. (1979). *Exploratory Data Analysis*. Beverly Hills, CA: Sage Publications.

Helbig, N., Cresswell, A. M., Burke, G. B., & Luna-Reyes, L. (2012). *The Dynamics of Opening Government Data: A White Paper.* Albany: Center for Technology in Government.

Huijboom, N., & Van den Broek, T. (2011). Open data: an international comparison of strategies. *European journal of ePractice*, *12*(1), 1-13.

Jetzek, T., Avital, M., & Bjørn-Andersen, N. (2012). The Value of Open Government Data: A Strategic Analysis Framework. In *2012 Pre-ICIS Workshop*.

Lakomaa, E., & Kallberg, J. (2013). Open Data as a Foundation for Innovation-The Enabling Effect of Free Public Sector Information for Entrepreneurs. *Access, IEE.* (1), 558-563.

Meijer, A.J., Curtin, D., Hillebrandt, M. (2012) Open government: connecting vision and voice. *International Review of Administrative Sciences*. 78(1), 10–29 .

Open Knowledge Foundation (OKF). (2012). *Open Data Handbook Documentation, Release 1.0.0.* Retrieved from: http://opendatahandbook.org/pdf/OpenDataHandbook.pdf

Open Government Licence Consultation Report. (2013, June 8). *data.gc.ca* Retrieved from: http://data.gc.ca/eng/open-government-licence-consultation-report

Openness Rating. (2013) *data.gc.ca.* Retrieved from: http://data.gc.ca/eng/openness-rating

Peristeras, V., Mentzas, G., Tarabanis, K. A., & Abecker, A. (2009). Transforming E- government and E-participation through IT. *Intelligent Systems, IEEE.* 24(5), 14-19.

Sawicki, D. S., & Craig, W. J. (1996). The democratization of data: Bridging the gap for community groups. *Journal of the American Planning Association*, *62*(4), 512-523.

Scassa, T. (2013, June 18). Canada's Open Government Licence V2.0 Is Released. Retrieved from: http://www.teresascassa.ca/index.php?option=com_k2&view=item&id=131:canada%E2%80%99s-open-government-licence-v20-is-released&Itemid=81

Shakespeare, S. (2013). Shakespeare Review: An independent review of public sector information. Department for Business, Innovation and Skills. UK

Ubaldi, B. (2013). *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives*. OECD Working Papers on Public Governance, No. 22, OECD

Uhlir, P. F., & Schröder, P. (2007). Open data for global science. *Data Science Journal*, *6*(0), 36-53.

## Appendix 1. Descriptive Statistics of Data Download Counts

**Case Processing Summary**

| | Cases | | | | | |
|---|---|---|---|---|---|---|
| | Valid | | Missing | | Total | |
| | N | Percent | N | Percent | N | Percent |
| Total unique downloads | 37 | 100.0% | 0 | 0.0% | 37 | 100.0% |

**Descriptives**

| | | | Statistic | Std. Error |
|---|---|---|---|---|
| Total unique downloads | Mean | | 79.41 | 24.963 |
| | 95% Confidence Interval for Mean | Lower Bound | 28.78 | |
| | | Upper Bound | 130.03 | |
| | 5% Trimmed Mean | | 54.60 | |
| | Median | | 19.00 | |
| | Variance | | 23056.026 | |
| | Std. Deviation | | 151.842 | |
| | Minimum | | 3 | |
| | Maximum | | 782 | |
| | Range | | 779 | |
| | Interquartile Range | | 74 | |
| | Skewness | | 3.371 | .388 |
| | Kurtosis | | 12.956 | .759 |

## Appendix 2. Frequencies of Values in Variables of Supply-Side Criteria for Effective Use

**Time period**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | since 2012 | 12 | 32.4 | 35.3 | 35.3 |
| | since 2008 | 22 | 59.5 | 64.7 | 100.0 |
| | Total | 34 | 91.9 | 100.0 | |
| Missing | 999 | 3 | 8.1 | | |
| Total | | 37 | 100.0 | | |

**Update interval**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Quarterly | 12 | 32.4 | 32.4 | 32.4 |
| | Yearly | 22 | 59.5 | 59.5 | 91.9 |
| | Other | 3 | 8.1 | 8.1 | 100.0 |
| | Total | 37 | 100.0 | 100.0 | |

**Info provided on acquisition**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Provided | 37 | 100.0 | 100.0 | 100.0 |

**Info provided on typical use**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Unexplained | 37 | 100.0 | 100.0 | 100.0 |

**Definitions of terms provided**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Undefined | 10 | 27.0 | 27.0 | 27.0 |
| | Simple | 23 | 62.2 | 62.2 | 89.2 |
| | Defined | 4 | 10.8 | 10.8 | 100.0 |
| | Total | 37 | 100.0 | 100.0 | |

**Accurate metadata**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| | Inaccurate | 25 | 67.6 | 67.6 | 67.6 |
| Valid | Accurate | 12 | 32.4 | 32.4 | 100.0 |
| | Total | 37 | 100.0 | 100.0 | |

**Tags correctly formatted**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| | Mistake | 22 | 59.5 | 59.5 | 59.5 |
| Valid | Correct | 15 | 40.5 | 40.5 | 100.0 |
| | Total | 37 | 100.0 | 100.0 | |

## Appendix 3. Cross-tabulations

### Demand for dataset * Time period

**Crosstab**

| | | | Time period | | Total |
|---|---|---|---|---|---|
| | | | since 2012 | since 2008 | |
| Demand for dataset | Low | Count | 6 | 13 | 19 |
| | | % within Time period | 50.0% | 59.1% | 55.9% |
| | High | Count | 6 | 9 | 15 |
| | | % within Time period | 50.0% | 40.9% | 44.1% |
| Total | | Count | 12 | 22 | 34 |
| | | % within Time period | 100.0% | 100.0% | 100.0% |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | .260[a] | 1 | .610 | | |
| Continuity Correction[b] | .022 | 1 | .882 | | |
| Likelihood Ratio | .260 | 1 | .610 | | |
| Fisher's Exact Test | | | | .724 | .439 |
| Linear-by-Linear Association | .253 | 1 | .615 | | |
| N of Valid Cases | 34 | | | | |

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.29.

b. Computed only for a 2x2 table

**Symmetric Measures**

| | | Value | Approx. Sig. |
|---|---|---|---|
| Nominal by Nominal | Phi | -.087 | .610 |
| | Cramer's V | .087 | .610 |
| N of Valid Cases | | 34 | |

## Demand for dataset * Update interval

**Crosstab**

| | | | Update interval | | | Total |
|---|---|---|---|---|---|---|
| | | | Quarterly | Yearly | Other | |
| Demand for dataset | Low | Count | 6 | 13 | 0 | 19 |
| | | % within Update interval | 50.0% | 59.1% | 0.0% | 51.4% |
| | High | Count | 6 | 9 | 3 | 18 |
| | | % within Update interval | 50.0% | 40.9% | 100.0% | 48.6% |
| Total | | Count | 12 | 22 | 3 | 37 |
| | | % within Update interval | 100.0% | 100.0% | 100.0% | 100.0% |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 3.703[a] | 2 | .157 |
| Likelihood Ratio | 4.863 | 2 | .088 |
| Linear-by-Linear Association | .578 | 1 | .447 |
| N of Valid Cases | 37 | | |

a. 2 cells (33.3%) have expected count less than 5. The minimum expected count is 1.46.

**Symmetric Measures**

| | | Value | Approx. Sig. |
|---|---|---|---|
| Nominal by Nominal | Phi | .316 | .157 |
| | Cramer's V | .316 | .157 |
| N of Valid Cases | | 37 | |

## Demand for dataset * Definitions of terms provided

**Crosstab**

| | | | Definitions of terms provided | | | Total |
|---|---|---|---|---|---|---|
| | | | Undefined | Simple | Defined | |
| Demand for dataset | Low | Count | 2 | 17 | 0 | 19 |
| | | % within Definitions of terms provided | 20.0% | 73.9% | 0.0% | 51.4% |
| | High | Count | 8 | 6 | 4 | 18 |
| | | % within Definitions of terms provided | 80.0% | 26.1% | 100.0% | 48.6% |
| Total | | Count | 10 | 23 | 4 | 37 |
| | | % within Definitions of terms provided | 100.0% | 100.0% | 100.0% | 100.0% |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 12.843[a] | 2 | .002 |
| Likelihood Ratio | 14.855 | 2 | .001 |
| Linear-by-Linear Association | .349 | 1 | .554 |
| N of Valid Cases | 37 | | |

a. 3 cells (50.0%) have expected count less than 5. The minimum expected count is 1.95.

**Symmetric Measures**

| | | Value | Approx. Sig. |
|---|---|---|---|
| Nominal by Nominal | Phi | .589 | .002 |
| | Cramer's V | .589 | .002 |
| N of Valid Cases | | 37 | |

## Demand for dataset * Accurate metadata

**Crosstab**

| | | | Accurate metadata | | Total |
|---|---|---|---|---|---|
| | | | Inaccurate | Accurate | |
| Demand for dataset | Low | Count | 13 | 6 | 19 |
| | | % within Accurate metadata | 52.0% | 50.0% | 51.4% |
| | High | Count | 12 | 6 | 18 |
| | | % within Accurate metadata | 48.0% | 50.0% | 48.6% |
| Total | | Count | 25 | 12 | 37 |
| | | % within Accurate metadata | 100.0% | 100.0% | 100.0% |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | .013[a] | 1 | .909 | | |
| Continuity Correction[b] | .000 | 1 | 1.000 | | |
| Likelihood Ratio | .013 | 1 | .909 | | |
| Fisher's Exact Test | | | | 1.000 | .593 |
| Linear-by-Linear Association | .013 | 1 | .911 | | |
| N of Valid Cases | 37 | | | | |

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.84.

b. Computed only for a 2x2 table

**Symmetric Measures**

| | | Value | Approx. Sig. |
|---|---|---|---|
| Nominal by Nominal | Phi | .019 | .909 |
| | Cramer's V | .019 | .909 |
| N of Valid Cases | | 37 | |

## Demand for dataset * Tags correctly formatted

**Crosstab**

| | | | Tags correctly formatted | | Total |
|---|---|---|---|---|---|
| | | | Mistake | Correct | |
| Demand for dataset | Low | Count | 15 | 4 | 19 |
| | | % within Tags correctly formatted | 68.2% | 26.7% | 51.4% |
| | High | Count | 7 | 11 | 18 |
| | | % within Tags correctly formatted | 31.8% | 73.3% | 48.6% |
| Total | | Count | 22 | 15 | 37 |
| | | % within Tags correctly formatted | 100.0% | 100.0% | 100.0% |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 6.153[a] | 1 | .013 | | |
| Continuity Correction[b] | 4.604 | 1 | .032 | | |
| Likelihood Ratio | 6.347 | 1 | .012 | | |
| Fisher's Exact Test | | | | .020 | .015 |
| Linear-by-Linear Association | 5.987 | 1 | .014 | | |
| N of Valid Cases | 37 | | | | |

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.30.

b. Computed only for a 2x2 table

**Symmetric Measures**

| | | Value | Approx. Sig. |
|---|---|---|---|
| Nominal by Nominal | Phi | .408 | .013 |
| | Cramer's V | .408 | .013 |
| N of Valid Cases | | 37 | |

**Appendix 4. Metadata**

**Contact information:** Cody Skinner, skinnerc@uoguelph.ca

All data was sourced from Canada's federal open data portal at data.gc.ca

**Title of datasets included in the sample (n = 37)**:
Applications Received at Case Processing Region (CPC-PRC) - Permanent Resident Card (in Persons)
Business Line - Applications Received for Permanent Residents (in Persons)
Business Line - Applications Received for Temporary Residents (in Persons)
Business Line - Authorizations and Visas Issued for Permanent Residents (in Persons)
Business Line - Visas, Permits and Extensions Issued for Temporary Residents (in Persons)
Canada - Permanent and Temporary Residents
Canada - Permanent residents by category
Canada - Permanent residents by province or territory and urban area
Canada - Permanent residents by source country
Canada - Total entries of foreign students by gender and level of study
Canada - Total entries of foreign students by province or territory and urban area
Canada - Total entries of foreign students by source country
Canada - Total entries of foreign workers by gender and occupational skill level
Canada - Total entries of foreign workers by province or territory and urban area
Canada - Total entries of foreign workers by source country
Canada - Total entries of humanitarian population by gender and age
Canada - Total entries of humanitarian population by province or territory and urban area
Canada - Total entries of humanitarian population by source country
CIC Operational Network at a Glance
Citizenship
Inventory - Permanent Resident Card (in Persons)
Permanent Resident Applicants Awaiting a Decision
Permanent Resident Applications Processed Abroad and Processing Times
Permanent Resident Cards Produced by Canadian Bank Note (CBN)
Permanent Resident Inventory
Permanent Resident Summary by Mission
Permanent Resident Visa Applications Received Abroad
Point of Service - Applications Received for Permanent Residents (in Persons)
Point of Service - Applications Received for Temporary Residents (in Persons)
Province Territory of Declared Destination - Authorizations and Visas Issued for Permanent Residents (in Persons)
Temporary Residents Applications Processed Abroad and Processing Times
Top 10 Source Countries - Applications Received for Permanent Residents (in Persons)
Top 10 Source Countries - Applications Received for Temporary Residents (in Persons)
Top 10 Source Countries - Authorizations and Visas Issued for Permanent Residents (in Persons)
Top 10 Source Countries - New Canadian Citizens (in Persons)
Top 10 Source Countries - Refugee Claims at All Offices (in Persons)
Top 10 Source Countries - Visas, Permits and Extensions Issued for Temporary Residents (in Persons)

**Variables:** (Missing data were assigned values of 999, or in the case of dates 09/09/9999)

| Name | Label | Type | Values |
|---|---|---|---|
| name | Dataset name | nominal | string |
| filename | Filename | nominal | string |
| desc | Description | nominal | string |
| format | File format | nominal | string |
| licence | Data licence | nominal | string |
| language | Languages offered | nominal | Numerical<br>0 = English<br>1 = French<br>2 = Both English and French |
| published | Date published | interval | date |
| modified | Date last modified | interval | date |
| cvgSTART | Starting date of coverage | interval | date |
| cvgEND | Ending date of coverage | interval | date |
| cvgMONTHS | Time period | nominal | Numerical<br>18 = Since 2012<br>66 = Since 2008 |
| update | Update interval | nominal | Numerical<br>0 = Quarterly<br>1 = Yearly |
| rtgTBL | Open Rating | ordinal | Numerical |
| rtgsAVGSCR | Average user rating | ordinal | Numerical |
| rtgsNUM | Number of user ratings | Interval | Numerical |
| cmtsNUM | Number of comments | Interval | Numerical |
| tags | Tags | nominal | string |
| publisher | Publishing Agency | nominal | string |
| uniqueDL_TOTAL | Total unique downloads | Interval | Numerical |
| uniqueDL_EN | Downloads in English | Interval | Numerical |
| uniqueDL_FR | Downloads in French | nominal | Numerical |
| downloadHiLo | Demand for dataset | Ordinal | Numerical<br>0 = Low<br>1 = High |
| collection | Info provided on acquisition | nominal | Numerical<br>0 = not provided<br>1 = provided |
| use | Info provided on typical use | nominal | Numerical<br>0 = unexplained<br>1 = explained |
| definitions | Definitions of terms provided | nominal | Numerical<br>0 = undefined<br>1 = simple<br>2 = defined |
| metadata | Accurate metadata | nominal | Numerical<br>0 = inaccurate<br>1 = accurate |
| tag | Tags correctly formatted | nominal | Numerical<br>0 = mistakes<br>1 =correct |