Spatial Data Standardization for Rural Open Data

Eric Hoodicoff

Geographic Information Systems, Selkirk College, erichoodicoff@edu.selkirk.ca

Abstract

The "Open Data for Open Government in Rural BC" project was approved in early 2016 for the Selkirk Geospatial Research Centre (SGRC) by The Social Sciences and Humanities Research Council (SSHRC 2016). This 3-year project involves the SGRC partnering with the Rural Development Institute (RDI), various Geographic Information Systems (GIS) professionals with varying backgrounds and open data experts to provide policy guidelines and recommendations for geospatial data from various rural communities and regional districts in the Kootenay region of British Columbia. To fully experience the benefits of open government data, the data must be interoperable and therefore, one of the objectives of this project is standardization. Through literature review, interviews with key stakeholders, research of other similar examples (including questionnaires sent to people involved in prior data standardization projects), and review of the data in question, I propose to create and document the process for standardizing the geospatial data involved in this project and possibly for other similar projects in the future.

Background and Relevance

From a conceptual point of view, data can be seen as the lowest level of abstraction from which information and then knowledge are derived (Ubaldi 2013, p. 5). Geospatial data are data linked to spatial coordinates. Therefore, the data represent actual areas on the Earth. Common municipal geospatial datasets are zoning areas, cadastral polygons and official community plans. Open Data are data that can be freely used, re-used and distributed by anyone, only subject to (at the most) the requirement that users attribute the data and that they make their work available to be shared as well (Ubaldi 2013). The rise of open data in the public sector has sparked innovation, driven efficiency, and fueled economic development. And in the vein of high-profile federal initiatives like Data.gov and the White House's Open Government Initiative, more and more governments at the local level are making their foray into the field with Chief Data Officers, open data policies, and open data catalogues (Goldstein and Dyson 2013).

The main goal of the "Open Data for open Government in Rural BC" project (I will refer to it as the "Open Data Project" for the remainder of this paper) is to investigate the potential outcomes of open data policy at the local scale for rural regions and communities through a detailed case study situated in British Columbia's rural Kootenay region. (Parfitt 2015). The Kootenay region is situated in The Canadian Columbia Basin in southeastern British Columbia (See the map in Figure 1). There are 28 communities in The Columbia Basin. Also 4 regional districts are either fully or partly situated there. These 32 separate local government entities all collect and use spatial data in some form. If all of these data were to be made open, there could be significant benefits for the municipalities, regional districts and the public. Benefits of open data include transparency, innovation, improved government efficiency and

effectiveness, and public participation. The focus of the open data project is geospatial data but all data will be considered. Planned outcomes for the project include policy direction for regional and local rural governments, data sharing and standardization agreements, and costed open data delivery options (Parfitt 2015).

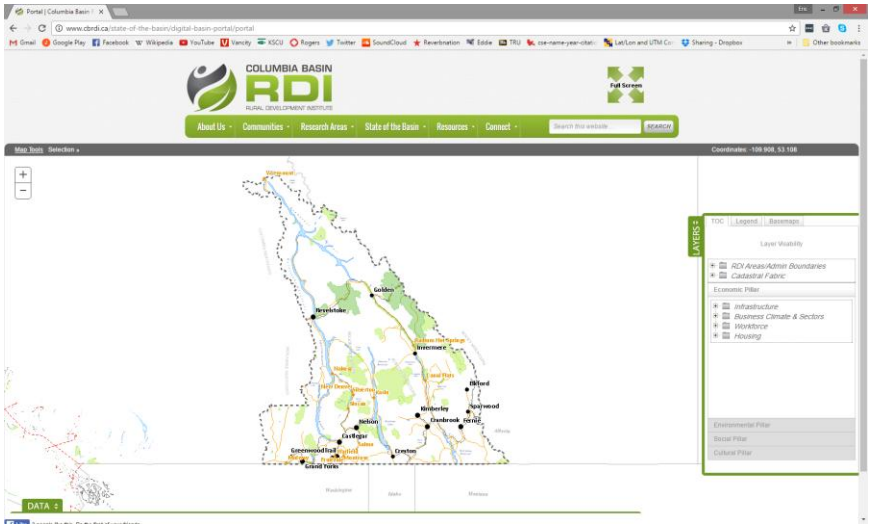

Figure 1: The Digital Basin Website showing the full extent of the Canadian Columbia Basin

Many larger municipalities around the globe have had much success with making their data open (two common examples being Amsterdam (Bicknese and van der Oord 2015) and Helsinki (Jaakola et al .2015)), but I haven't been able to find many examples of smaller rural communities doing this. Open data is a relatively new trend and there might not be very many rural areas opening their data yet.

I spent the summer of 2016 working with the actual spatial data delivered by the 32 government entities for a previous SGRC project called "The Digital Basin Portal" (See figure 1) which can be looked at as the precursor to the Open Data Project. This web mapping application portrays over 200 indicators of well being and was developed by the SGRC partnering with the RDI. Data sharing agreements were developed and Geospatial data were collected from the communities and regional districts in the Columbia Basin and were made available for public viewing on a web application (Ector, 2015). The portal can be viewed online at http://www.cbrdi.ca/state-of-the-basin/digital-basin-portal/. One of my roles in the project was to assess the quality of the data that had been delivered to the RDI for the portal.

While I worked with the data, I noticed a few things. The larger the city, the more data were commonly available. Often, the smaller municipalities had the regional district take ownership of their spatial data. Most importantly, I noticed similarities in the types of data (cadastral, official community plans, etc.) but the data lacked consistency. The schemas were different. Attribute tables were set up and filled out differently. Definitions differed (i.e. what some communities defined as official community plans were different than others and this can be seen on the digital basin portal). Also, metadata was almost non-existent except for a small few cases. Merging the data from the different communities together would be difficult as each entity designed their

databases and schemas to work for them and how they operate. One could merge the data by converting file types, schemas and cleaning up the data but doing so would be very time-consuming.

According to Sieber and Johnson (2015), municipal governments hold numerous goals in publishing data, among them, to encourage local economic development, improve service delivery, provide internal business intelligence, and increase transparency and public accountability (as cited in Bloom & Sieber, 2016). A precondition for any of these goals is datasets that are machine readable and interoperable. One way to ensure machine readability and interoperability are through open data standards.
Open data standards provide the semantic and schematic guidelines to ensure machine readability and interoperability so governments can actually succeed in opening up data to the public (Bloom & Sieber, 2016). Data alone is of little use unless those data are presented in a form that is amenable to analysis and interpretation (Brooksbank and Quackenbush 2006). Aalders and Hunter (2009) explain how standardization applies to GIS and geospatial data:

> With the evolution of the "information age" over the past 45 years, standards have also been required for the effective and efficient dissemination and application of information and information technology – and the needs of spatial information have been no exception. At first, standards were developed that would enable GIS technology to function correctly – for example in user interfaces, computer operating systems, database query languages, and network communications – but generally these standards did not apply directly to spatial data. However, this soon changed as users sought to share the digital spatial datasets they had developed in order to reduce data duplication, minimize data collection costs and implement integrated data applications.

In order for consumers and third parties to process information cheaply and efficiently, such information should be available in standardized vocabularies and formats which allow for meaningful comparisons and other analyses across datasets (Sunstein 2011). One of the goals of the Open Data Project is standardization for these reasons. With this project, I won't be necessarily doing the standardization itself. Instead, and under the guidance of the head of the Open Data Project (and my supervising instructor for this thesis), I propose to create and document the best process for standardization of the Open Data Project and possibly other similar projects in the future.

What makes this project unique is the rural context. There are many examples of urban open data but not many examples of rural open data, especially in North America. There are a few papers on challenges with rural open data in developing countries (for instance Gurstein, 2011 and Schwegmann, 2012). However, most of the challenges mentioned deal with much of the population lacking computers and applicable skills and do not really apply in our situation. Although our broadband access lags behind what is available in larger cities, there is no indication that access to personal computers or computer skills is lower in our region. Further research might even show that the challenges our rural area face might even differ from other rural areas in Canada.

One significant unique difference in our rural data versus urban data, like Vancouver for example, is the size of the area. Vancouver is one city while our rural data is spread over 28 communities and 4 regional districts. Driving from one side of Vancouver to the other on a good day would take less than an hour. Driving from the southern-most city of Trail to the northernmost community of Valemount would take almost 9 hours. The terrain, culture and history differ throughout the Columbia Basin which can affect the data that describes the area thus adding to the challenge of proving interoperable data. Also, it will be a challenge to get all of the stakeholders and partners in one room to discuss the project.

Another significant difference is the city of Vancouver is home to over 603,000 people while the Columbia Basin is home to only roughly 150,000. We have fewer tax dollars going to our data collection and management in the Columbia Basin. The datasets deal with fewer people and less complex infrastructure but a larger area spread over many entities. The data isn't as complex and the software used and file formats are not always as current as a larger city's would be. The data from one community could be very different than another based on the area surrounding the community, the size of the community, the size of the population, and the current and historic methods of data collection.

Finally, I'd like to mention that some of the smaller communities (for instance New Denver with a population of around 500) only have 3 or 4 employees working in the town office. Open data and GIS are brand new to these communities. On the other end of the spectrum, some larger communities (for instance Cranbrook with a population of over 19,000) have been using GIS for years now, have an extensive up-to-date catalogue of data and imagery available and are fully aware of the benefits of open data. There is an opportunity here to provide standards and examples for the smaller communities to adapt for not only opening their data but collecting the data itself.

Methods and Data

Creating a standard is hard. The right way to create a standard involves engaging a broad range of stakeholders in the public and private sectors, including producers and consumers of data in that format, to create something that will be broadly useful and stand the test of time. (Jaquith 2016)

The potential communities involved are:

City of Castlegar
City of Cranbrook
City of Fernie
City of Grand Forks
City of Greenwood
City of Kimberley
City of Nelson
City of Revelstoke
City of Rossland
City of Trail
District of Elkford
District of Invermere
District of Sparwood
Regional District of Central Kootenay
Regional District of East Kootenay
Regional District of Kootenay Boundary
Town of Creston
Town of Golden
Village of Radium Hot Springs
Village of Canal Flats
Village of Fruitvale
Village of Kaslo
Village of Midway
Village of Montrose
Village of Nakusp
Village of New Denver
Village of Salmo
Village of Silverton
Village of Slocan
Village of Valemount
Village of Warfield
Village of Kaslo

The first step of the project will be a literature review of existing standards and standards processes. Even larger municipalities' standardization processes can be useful. What works? What doesn't work?

With this information, I will interview (in person and with questionnaires) the partners of the Open Data Project as well as the potential stakeholders in the various government offices in these communities. What standards are already in place? What lessons have they learned from their current standardization policies and can they offer any advice? What would they like to see in the standardization process for this project?

I have also spent a significant amount of time assessing the quality of the data delivered to the RDI and SGRC for the Digital Basin Portal project. I have personally seen what the data looks like, how the attributes are filled out, what metadata exists, etc. I will continue to look at this data while working with the GIS professionals at the various communities to assess what works and what doesn't work and to hopefully come up with a solution that works for all involved.

One note on having multiple stakeholders involved and pushing the project forward is this: There is a line that divides full consensus among the stakeholders and eventually simply just choosing 1 or 2 stakeholders and creating the standards then asking the others to follow suit. This project might end up following this same lead in the end if consensus seems too difficult.

Jaquith (2016) explains how reaching a consensus could be challenging and how the General Transit Feed Specification (GTFS) were successful by "just doing it".

> GTFS is the huge success story here, and that resulted from some Google engineers working with a single transit agency. There was no series of roundtables, no acceptance testing, no RFC. They just did it, building something lightweight and extensible that solved the problems at hand. It's changed a lot in the 11 years since, adapting to the needs of its growing user base and becoming subject to the normal standards-creation processes, but for almost that entire time, GTFS has been *the* standard for transit data.
> What we need is for tiny groups of stakeholders—maybe mere pairings of stakeholders—to *just go ahead and create standards within their area of expertise*. And don't call it a "standard," if that sounds too scary. Call it an "implementation" or "our schema," or whatever. Develop it in the open, document it, set up a validator, put it to work, and get out the word that it exists. (Jaquith, 2016)

With all of this information, I will create a first draft of the standardization process and share it with the stakeholders for feedback. I will then revise the process from that feedback.

Results

The result of this project will be a process of standardization with the following fundamental aims as described by Aalders and Hunter (2009, p.5):

- Efficiency (standards lead to improved data sharing due to easier data transfer, thereby avoiding costly and time-consuming duplication of data collection and processing);
- Avoidance of Information Loss (the use of common standards help minimize the loss of data that usually occurs when transforming data from one system to another);
- Portability of Applications (often specialized application software will be developed for spatial data, which should be able to be shared by users of different software and hardware platforms);
- Ease of Learning and Increased Productivity (shared application software means that other users can benefit from using those applications without having to develop their own software, thus saving time and money);
- Quality Improvement (standards make it possible to provide clear and well-defined quality concepts); and
- Knowledge Transfer (standards help clarify aspects of data usage and help different users transferring spatial data from one to another system to better understand the requirements of other users)

All of the above mentioned aims apply to this project. Smaller communities with little data experience will benefit from the available standards, templates, examples and open discussions with other communities. The larger, more experienced communities will benefit with better efficiency and improved quality. The public will benefit with interoperable data that can be useful for business and personal decision-making. The local governments will have more transparency thus increasing trust and decreasing the amount of work to answer questions that this data will answer for them.

Conclusions

I am happy to be a part of this Open Data Project and I am looking forward to contributing to it. It will be interesting to gather information on how other organizations standardize their open data and to see how it will work in the rural British Columbia setting. I also look forward to speaking further with the project partners and potential stakeholders. The benefits of opening up the spatial data for these communities could be very significant and I am excited to see how the entire project turns out.

References

Aalders H.J.G.L. and Hunter G.J. (2009). Spatial data standards. *Advanced geographic information systems*. Encyclopedia of Life Support Systems

Bicknese L and Van der Oord M. (2015). Open city statistics: The first results with open data in Amsterdam. *Statistical journal of the IAOS*. 31: 111-115

Bloom R and Sieber R. (2016). Open civic data standards in Canada. Geothink. 1

Brooksbank C. and Quackenbush J. (2006). Data standards: A call to action. *OMICS: A journal of integrative biology*. Mary Ann Liebert Inc. 10(2), 94-99

Community and college social innovation fund (CCSIF) November 2015 competition awards [Internet]. (2016). Social Sciences and Humanities Research Council; [updated 2016 Jun 21: cited 2016 Nov 9]. Available from: http://www.sshrc-crsh.gc.ca/results-resultats/recipients-recipiendaires/2015/ccsif-fiscc-eng.aspx

Digital Basin Portal [Internet]. (2016). Castlegar, BC. Selkirk Geospatial Research Centre and Rural Development Institute. [Cited on 2016 Oct 10]. Available from http://www.cbrdi.ca/state-of-the-basin/digital-basin-portal/

Ector S. (2014). Digital basin: Mapping the state of well-being in rural British Columbia. Castlegar, BC. Selkirk Geospatial Research Centre

Goldstein B. and Dyson L. (2013). *Beyond transparency: Open data and the future of civic innovation*. Code for America Press.

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

Jaakola A, Kekkonen H, Lahti T and Manninen. (2015). Open data, open cities: Experiences from the Helsinki metropolitan area. Case Helsinki Region Infoshare www.hri.fi. *Statistical Journal of the IAOS*. 31: 117-122

Jaquith W. (2016). We Need Data Schemas – So Let's Create Them. [Internet] U.S. Open Data [Cited on 2016 Oct 12]. Available from: https://usopendata.org/2016/07/29/schemas/

Parfitt, I. (2015). *Open data for open government in rural BC*. Castlegar, BC. Selkirk Geospatial Research Centre. 4-5

Schwegmann C. (2012). Open Data in Developing Countries. *EPSI Platform 2013:02*

Sieber R & Johnson P. (2015). Civic open data at a crossroads: Dominant models and current challenges. *Government Information Quarterly* 32(3): 308-315

Sunstein C.R. (2011). Informing consumers through smart disclosure. *Memorandum for the heads of executive departments and agencies*. 5

Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives. *OECD working papers on public governance, No. 22*. OECD Publishing. 5-6