GEGN.2/2019/CRP.133

15 March 2019

Original: French

United Nations Group of Experts On Geographical Names
First session
New York, 29 April – 3 May 2019
Item 15 of the provisional agenda
Toponymic data files and gazetteers (data processing and tools, database management, data dissemination: products and services) (Working Group on Toponymic Data Files and Gazetteers)

Rationalizing and normalizing the status codes of geographical names

Submitted by Canada**

* GEGN.2/2019/1

** Prepared by Cindy Doyle, Canada Centre for Mapping and Earth Observation, Natural Resources Canada

Summary :

The status of a geographical name must indicate whether it is an official name or a rescinded historical name. However, although the Canadian Geographical Names Database (CGNDB) was completely rebuilt in 2015 according to a normalized spatial model, it was not possible at that time, to eliminate the implicit information redundancy hidden in the status codes. Redundancy is when information is stored in two different places in the database, which can lead to inconsistencies and even contradictions if the information is subsequently updated in only one of those places. The sheer number of place-name status codes in the CGNDB was the most patent example, further complicated by the use of jargon that only an expert could decode. The Geographic Names Board of Canada (GNBC) Secretariat addressed this issue by establishing a committee tasked with redefining the list of status codes to make them clear, concise, interoperable and free of any redundancy that could lead to contradictions. This article describes the work done by Natural Resources Canada (NRCan) to rationalize the former, cumbersome and outdated list of codes to make it simple and intuitive, while facilitating exchanges between the CGNDB and provincial partners.

Background

The Geographical Names Board of Canada (GNBC) is the national coordinating body responsible for standards and policies for geographical naming in Canada. The Board is established under a federal Order in Council, and is composed of members from federal, provincial and territorial government departments and agencies, each with specific responsibilities for their respective jurisdictions and mandates. Working together as a multi-jurisdictional national body, GNBC members ensure that geographical names are consistently managed in Canada.

The GNBC is supported by a Secretariat provided by Natural Resources Canada (NRCan), a department of the Government of Canada. NRCan provides infrastructure and support for the Canadian Geographical Names Data Base (CGNDB), the national database of authoritative geographical names and a key component of Canada’s Spatial Data Infrastructure.

List of CGNDB status codes: before and after

It is said that a picture is worth a thousand words. One has only to look at the two lists in Table 1 — the old list, comprising 52 indecipherable codes, and the new one, with only 6 — to see how complex geographical name classification had become.

***Table 1* – Old and new lists of status codes**

Old status code list A1, A10, A11, A12, A13, A14, A2, A3, A4, A5, A6, A7, A8, A9, A99, B1, B10, B11, B12, B13, B14, B15, B2, B3, B4, B5, B6, B9, C1, C10, C11, C2, C3, C4, C5, C7, C9, D48, D60, D62, E1, E10, E2, E3, E4, E5, E6, E7, G1, M1, M2, P1, Q1, Q2, S1, S2, S3, T1, U1, U2, U3, W1, X1, Y1, Y2
New status code list - official - previouslyOfficial - pending - alternate - unpublished - forInformation

Without a data dictionary to decipher the status codes, it was impossible for a lay person to determine, from a status code in the old list, whether a geographical name was official or historical.

How had we gotten there?

Origin of the CGNDB status codes

To understand where this jargon comes from, we have to go back to the days when memory bytes were a rare and expensive commodity. When databases were assembled at that time, anything that could be shortened to save space was shortened. The most common example was to use only the last two digits of the year to indicate a date, which led to the well-known Y2K bug. Also in the interest of space-saving, most of the attributes of a geographical name were coded to shorten the information they contained and make them "digestible" for the limited capacity of the computers of the time. However, this abbreviation of the status codes does not explain the existence of 52 variants. Since the CGNDB is a mosaic of data from ten provinces, three territories and four federal agencies, each with its own independent system, the status codes were not necessarily harmonized when imported, resulting in multiple code variants with the same meaning. In addition, the early versions of the application used to update the CGNDB were not easily adaptable or programmable in order to meet new management needs. The CGNDB secretariat used status codes by assigning fictitious codes for the sole purpose of grouping them together and easily retrieving them from the database. For example, in order to be able to find a dataset corresponding to a batch of names uploaded on a given date, an attribute then unavailable for searching purposes, a fictitious status code such as D48 was given: D for *digital content* and 48 for *Alberta*. This code allowed toponymists to find all names uploaded on a given date by using the status code D48 in the search criteria. These manipulated codes, although relevant and useful at the time, had no relationship to the actual status of the geographical name and distorted the attribute that was supposed to reveal whether the name was official or rescinded. These codes were generally used temporarily, but some were forgotten until the 2018 normalization wiped the slate of this bygone era.

Normalization

According to Wikipedia, *database normalization is the process of structuring a relational database in accordance with a series of so-called normal forms in order to reduce data redundancy and improve data integrity.*

There are several levels of normalization and, without going into detail, the attributes that conform to the first normal form are said to be atomic, that is, the subdivision of information does not provide any additional or complementary information. The old code list did not satisfy this criterion. In addition to indicating whether the name was official or not, it provided additional information on the origin or geographical location of the name, the nature of the feature named and even the use of the name, in blatant violation of the first level of normalization.

The second and third normal forms prohibit inter-attribute dependencies, that is, the value of one attribute must not duplicate the value contained in another attribute. The status codes in the old list, in addition to providing too much information, duplicated information contained in other attributes, such as province or feature type, creating dependency between them. These attributes frequently became inconsistent when only one of the two was subsequently updated.

For example, the status code for an official Indian reserve name was A13. It could happen that the Indian reserve lost its official reserve status but that the name was still in use by the community. A well-meaning toponymist would change the generic code of the Indian reserve to that of a community, but would neglect to change the status code A13, which should be for the exclusive use of Indian reserves. Two contradictory pieces of information defined the name: its generic code and its status code, which stipulated that it was a community and an Indian reserve, respectively. In such a case, the user was unable to determine which was valid—the generic code or the status code—unless he or she researched the decisions. This research sometimes required contacting the names authority to validate the information, which was a waste of time. This waste of resources was the result of poor attribute modelling, which allowed these inconsistencies to exist. Although the complete redesign of the CGNDB in 2015 required numerous consultations over a period of a year, this list survived, against all the rules of sound management of a relational database.

What the codes concealed

It might appear, from Table 2 below, that the status codes were organized and classified correctly. However, these categories overlap, such that a geographical name sometimes fell into four different categories, making classification a random exercise, for want of clear guidelines.

***Table 2* – Definition of the alphabetical portion of the status codes of the old list**

Code	Definition
A	Approved
B	Not approved
C	Never approved
D	Digital information
E	Quebec name
M	Manitoba name
P	Pan-Canadian name
Q	Ontario name
S	Statutory name
W	Approved international waters name
U	Used in special circumstances
Y	Yukon name

In which category is an approved Quebec geographical name classified? Or a rescinded Ontario name? Or a dual Manitoba name received in digital format? It was impossible to know unless you had the data dictionary providing the definition of each code and instructions for its use. This document consisted of 15 pages of tables and definitions, and required extensive training and experience in order to learn how to use them correctly.

Table 3 is a sample of the main status codes from the old list, with the definition as it appeared in the dictionary, the implicit information the codes contained, and the normalized replacement codes. Following this are the solutions provided to compensate for the information loss resulting from the reclassification of the old status codes to the six normalized codes.

***Table 3* – Main old status codes**

Code	Definition	Hidden or duplicated information contained in another attribute	Normalized code
A1	Approved		official
A3	Dual approval	Name approved by more than one authority	official
A7	Dual name	Feature for which there are two official names	official
A8	Name change	Name replaced by another name	official
A10	Official name in Quebec	Province of Quebec	official
A13	Indian reserve name	Feature type, Indian reserve	official
B2	Rescinded name	Name not replaced	previouslyOfficial
B3	Deleted name	Name published inadvertently	forInformation
B4	Rescinded name	Name replaced by a new official name	previouslyOfficial
B5	Rescinded Indian reserve name	Feature type, Indian reserve	unpublished
B10	Name rescinded in Quebec	Province of Quebec Name not replaced	previouslyOfficial
B11	Published cross-reference	Province of Quebec Name replaced	previouslyOfficial

Solutions for preventing information loss

In most cases, information associated with the status code was redundant, already included in other attributes of the name, and their removal did not result in information loss. The following two codes are examples in which no action was taken to save the hidden information, since it was redundant and potentially detrimental (Table 4).

***Table 4* – Redundant information that was not saved**

Code	Definition	Hidden or duplicated information contained in another attribute	Normalized code
A10	Official name in Quebec	Province of Quebec	official
A13	Indian reserve name	Feature type, Indian reserve	official

Code A10 designated official geographical names in Quebec. However, province information is already included in the REGION_CL attribute, which is a searchable, indexed field. No action was taken to retain this redundant information. It was simply replaced with the new official geographical name code officially approved by the names authority.

Similarly, Code A13 designated official Indian reserve names. Feature type is also included in a normalized list under *IR – Indian Reserve/Réserve Indienne.* Like the regions, the feature type is a searchable, indexed field. Users looking for the official name of an Indian reserve can find the information using a combination of two search criteria: status and feature type.

This is not the case for the examples presented in tables 6 and 7, where the information added was found to be non-redundant and was considered important enough to remain in the toponymic groups. Toponymic groups are collections of geographical names that make searching individual names easier. One of their greatest advantages is that they are completely automated: they are updated weekly by processes that analyze data and reconstitute the groups without manual intervention.

***Table 5* – Toponymic groups created to compensate for information loss when the status codes were rationalized**

	NAME	DESIGNATION_EN	
1244	jointDecision	Joint Decision	The *Joint Decision* group is used to identify names that require the approval of more than one authority, such as the provincial or territorial authority and the appropriate federal agency authority.
1245	multiple	Multiple	The *Multiple* group is used to identify geographical names for features with more than one officially approved name.
1247	rescinded	Rescinded	The *Rescinded* group is used to identify geographical names that have had their official recognition removed and that have not been replaced by a new name.

The second advantage is that they allow geographical names to belong to several groups. Whereas once, a single status code had to be chosen for a given geographical name, it can now belong to several groups. As a result, French and English names of national parks affected by a joint provincial-federal decision can be in both the *jointDecision* group and the *Multiple* group. Belonging to the *jointDecision* group indicates that two authorities issued a decision regarding the name, and belonging to the *Multiple* group indicates that the name listed is not the only official one, i.e. there is more than one name in more than one language. With the old code list, it was impossible to gather all of this information. The toponymist had to choose between Code A3 – *Dual Approval* or Code A7 – *Dual Name*, whereas in fact both definitions applied to the name (Table 6). Status code assignment was inconsistent and the information was incomplete.

***Table 6* – Information saved in the jointDecision and Multiple toponymic groups**

Code	Definition	Hidden information that was saved	Normalized code	Group retaining the information
A3	Dual approval	Geographical name approved by more than one authority	official	*jointDecision*
A7	Dual name	Feature with two official names	official	*multiple*

With normalization, the toponymist no longer has to worry about the number of approved names or the number of decisions issued by the various authorities, and can instead enter the decisions and names with their official statuses. The weekly processes will auto-populate the toponymic groups.

***Table 7* – Information saved in the rescinded group**

Code	Definition	Hidden information that was saved	Normalized code	Group retaining the information
B2	Rescinded name	Name not replaced	previouslyOfficial	*rescinded*
B4	Rescinded name	Name replaced by a new official name	previouslyOfficial	N/A
B10	Name rescinded in Quebec	Province of Quebec Name not replaced	previouslyOfficial	*rescinded*
B11	Published cross reference	Province of Quebec Name replaced	previouslyOfficial	N/A

All codes in Table 7 were assigned to historical names. Codes B2 or B10 (in Quebec) were used when rescinded names were not replaced by new names. Codes B4 and B11 (in Quebec) were used when the rescinded names were replaced by new names. However, a name could be replaced several years after it was rescinded. When a new name was added, the status of the historical name was not always changed to B4 or B11. When this happened, the name's status would continue to indicate that the name had not been replaced, which was not the case. The new code list simplifies these operations: a single code (*previouslyOfficial*) is assigned to a rescinded name regardless of whether it is located in Quebec or whether it has been replaced. The automated update process for toponymic groups adds the names that have not been replaced to the *rescinded* group. If, years later, new replacement names are approved, no action is required to indicate that the rescinded names are being replaced. The automated processes detect new names that share the feature's unique identifier and automatically remove the historical name from the *rescinded* group.

Conclusion

This normalization of status codes has allowed us to eliminate redundancies, potential inconsistency factors and, above all, significant time loss caused by the complexity of the jargon used, in terms of training employees and manually maintaining incomplete information. In addition, the new status codes are straightforward (Table 1), allowing the general public to easily understand them. We are proud to have established this consensus among GNBC members to produce this simple and easily understandable list of codes for anyone who uses Canada's geographical names.