Metadata and Controlled Vocabularies in the Government of Canada: A Situational Analysis

Gregory Renaud
Treasury Board of Canada, Secretariat
Government of Canada
Tel. 613-946-6280
Fax. 613-946-9342
Mail: renaud.gregory@tbs-sct.gc.ca

Abstract

This paper describes the Government of Canada's standards and recent activities to create and manage metadata and controlled vocabularies.

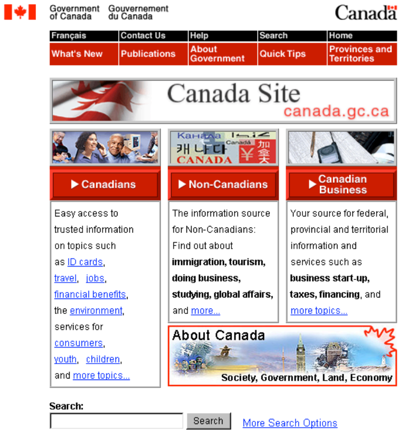
The Government of Canada (GoC) has been working actively for several years to enhance access to its published information through the use of metadata. In recognition of the value of controlled vocabularies in managing electronic information, the GoC has adopted standards for metadata and controlled vocabularies. Various initiatives have been proceeding to create and adopt controlled vocabularies for use with Dublin Core and other metadata schemas. Work is proceeding simultaneously on several fronts: establishing governance and developing tools to create and adapt controlled vocabularies, extensibility and interoperability frameworks, development of metadata registries and repositories, and creation and mapping of taxonomies. Canadian government departments and agencies have engaged in these metadata initiatives to support the fundamental priority of transforming services. The challenge is to allow the initiatives to mature and develop while ensuring they are co-ordinated.

Keywords
Canada, controlled vocabularies, metadata, metadata standard, interoperability

1. Background

In the 1999 Speech From the Throne, the Government of Canada promised to "…become a model user of information technology and the Internet." The Internet was recognized as an important channel for service delivery of citizen-centred services. The Government set the bar high: "…to be known around the world as the government most connected to its citizens, with Canadians able to access all government information and services on-line at the time and place of their choosing."[1] This commitment spawned the Government On-Line initiative, which led to a variety of projects to reinvent the way government services are offered to citizens and managed between organizations, with Internet technology as a primary enabler.

For the purpose of this paper, two of these initiatives are particularly



Easy access to trusted information on topics such as ID cards, travel, jobs, financial benefits, the environment, services for consumers, youth, children, and more topics...

The information source for Non-Canadians: Find out about **immigration, tourism, doing business, studying, global affairs,** and more...

Your source for federal, provincial and territorial information and services such as **business start-up, taxes, financing**, and more topics...

Fig. 1 Canada Site home page: Canadians, Non-Canadians and Business Gateways

https://doi.org/10.23106/dcmi.952107752

significant. First, the development of the "Common Look and Feel" standard, which set guidelines for the structure and appearance of all Government web sites. Included in this standard is the requirement for GoC departments and agencies to use metadata to describe creator, title, subject, date and language for web resources.[2]

Second, the management and structure of the Government of Canada web site laid the groundwork for interoperability of information. The main page of the Canada site offers "gateways" for three audiences: Canadians, non-Canadians and businesses. Each gateway contains sub-sites organized by audience, subject and activity. These sub-sites, called "clusters", are managed by teams from one or more federal departments. The gateways and clusters will soon be managing the content of their sites by collecting ("harvesting") metadata assigned to web resources by departments and agencies. In order for this to be successful, interoperability of metadata elements and vocabulary values are key requirements. This includes the understanding, tools and capacity to have them well and consistently applied.

In September 2000 a group of metadata experts from several federal departments met to develop and evaluate metadata standards for the GoC. This group became the Government On-Line Metadata Working Group, chaired by the Treasury Board of Canada, Secretariat. In 2001, on the recommendation of the Working Group, the Canadian Government adopted the Dublin Core as the core metadata standard for resource discovery.[3]

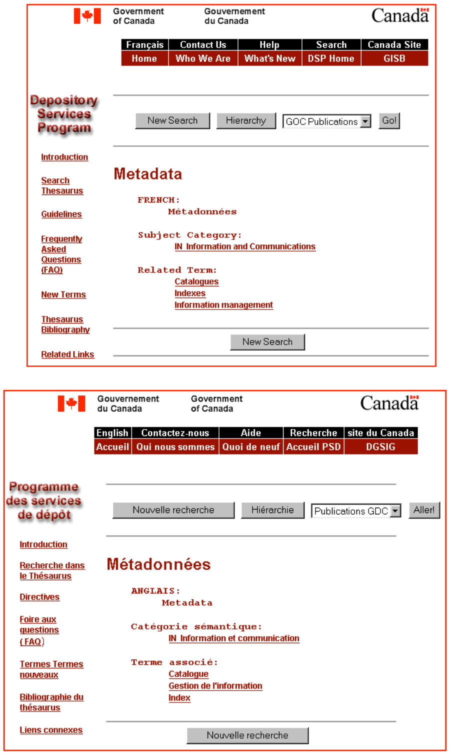
The Metadata Working Group has become the primary vehicle for developing a government-wide metadata strategy, management and policy framework. Through the work of its sub-groups, it has also steered the identification, creation and use of controlled vocabularies for resource discovery.

2. The Controlled Vocabulary Standard

The Controlled Vocabulary Standard, a corollary to the metadata standard, adopts the *principle* of controlled vocabulary for the management of electronic information.[4] It underscores the necessity of using controlled vocabularies to classify and describe information and to support navigation, searching, information sharing and interoperability goals of Government On-Line.

The Standard also adopts the Government of Canada Core Subject Thesaurus (CST)[5] as the default thesaurus to be used by federal organizations for the "subject" element within web resources. The genesis of the CST was the Depository Services Program Subject Thesaurus, a source of subject descriptors for bibliographic records made available to Canadian libraries. The Controlled Vocabulary Standard also allows the use of internationally accepted controlled vocabularies, such as the Library of Congress Subject Headings, Canadian Subject Headings or the Répertoire des vedettes-matières de l'Université Laval.

https://doi.org/10.23106/dcmi.952107752



Metadata

FRENCH:
Métadonnées

Subject Category:
IN Information and Communications

Related Term:
Catalogues
Indexes
Information management

Métadonnées

ANGLAIS:
Metadata

Catégorie sémantique:
IN Information et communication

Terme associé:
Catalogue
Gestion de l'information
Index

Fig. 2 Canada's Core Subject Thesaurus: web-based source of English and French subject terms

However, given the comprehensiveness and ease of access of the CST, it is the most frequently used source of subject terms in the GoC for web resources. In addition, the CST was developed in accordance with ISO 2788-1986 (monolingual thesauri) and ISO 5964-1985 (multilingual thesauri). All GoC web resources and accompanying metadata must be available in English and French, Canada's two official languages. Since the CST includes equivalent English and French language terms, it is possible for indexers working in either language to easily apply appropriate terms to both language versions of the resource.

Encoding schemes were subsequently identified for the other mandatory elements, creator, language and date (there being no need for a title scheme, since content for this element is generally found on the resource itself). The identification of controlled vocabularies to populate the remaining elements of the Dublin Core has been driven by the elements which federal departments, agencies, gateways and clusters consider to be of greatest need. The initial elements were type, audience, geographic coverage and format.

3. Developing GoC Controlled Vocabularies

Generally, when a need is identified for a controlled vocabulary, the Metadata Working Group establishes a sub-group to determine the appropriate course of action: adopt, adapt or create a vocabulary. The type and audience sub-groups determined that they would need to create vocabularies, since the level of specificity of existing vocabularies, such as DCMI Type, was not sufficient. Existing vocabularies have been adapted for geographic coverage and format. During the development of the type and audience vocabularies, some broad principles were agreed upon to govern the selection of terms:

- **High level:** Terms represent broad concepts that may be further expanded into detailed schemes.
- **Applicable:** Terms represent content types found on a significant number of Government of Canada Web sites, and/or are of substantial significance to Government of Canada programs/services.

https://doi.org/10.23106/dcmi.952107752

- **Recognizable:** Terms are understandable by implementers and indexers.
- **Unique**: No terms will be a synonym of an existing term.
- **Client-centric:** Concepts and terminology used in presentation layer should be tested with the public

These principles help ensure that controlled vocabularies meet the practical needs of indexers as well as providing direction to maintain the integrity of the vocabularies themselves. Indexers may be any creator of a Web page; they are a diverse and distributed group. When used properly, the principles guarantee metadata quality and basic interoperability by ensuring that the terms will be recognized by the GoC site search engine and are useable by the Gateways and Clusters for selecting content for their sites.

The type and audience vocabularies are published and maintained by the Treasury Board of Canada Secretariat. The final reports of the sub-groups which developed the vocabularies, as well as the vocabularies themselves, are accessible on the Treasury Board of Canada Secretariat web site.[6]

Two controlled vocabularies have been identified for the geographic coverage element (dc.coverage.spatial). Unlike the type and audience vocabularies, these were not created by a sub-group. Robust, well-structured and maintained vocabularies already existed in other federal agencies. Although these vocabularies were not created expressly for use in Dublin Core metadata, they function well as vocabularies for the coverage element. A description of the vocabularies and how they are to be used is available on the Treasury Board of Canada Secretariat web site.[7]

The Internet Assigned Numbers Authority (IANA) Multipurpose Internet Mail Extensions (MIME) Type list[8] is often cited as a vocabulary to populate the format element. However, the GoC format sub-group decided not to adopt the entire IANA list as a default vocabulary, nor did they create an entirely new vocabulary.

In order for metadata elements and vocabularies to be properly understood and applied within the GoC, it is generally preferred that guidance on why, when, how to apply them, the syntax of the values be obvious, and that they be easily accessible. Another essential component is that the management of the vocabularies be responsive to the needs of the GoC institutions. Although the IANA MIME type list contains a comprehensive list of file formats, it does not include any guidelines on how to express it's values within a metadata element. The process for updating the list in a timely way is not clear.

The format sub-group also noted that the MIME Types categories include some values which appear in the GoC type scheme, which could be confusing to implementers. The IANA list contains approximately 400 terms, most of which would not be used to describe web resources and therefore did not meet the principle of "applicability". Finally, there are some values which did not appear in the IANA MIME Type list which the GoC wanted to use (e.g.

https://doi.org/10.23106/dcmi.952107752

shockwave, flash, realaudio). So the GoC format vocabulary was based largely on terms from the IANA list, but with additional terms, and published along with guidelines for its use and a mechanism for adding new terms.[9]

4. The Interoperability Value Proposition

At this point it is worthwhile to step back and consider some of the drivers for interoperability in the GoC.

4.1 Obligation

When departments and agencies were required to add Dublin Core metadata to their public-facing web pages, many resisted the expenditure of resources to create metadata. Some argued that metadata had, at best, a questionable role in leading commercial search engines to their sites. With the emergence of the Canada site and the Gateways and Clusters, the benefits of using metadata became more focussed on search and retrieval within the government of Canada domain, over which it is possible to exert some control. Still, the discussion around applying metadata resulted in a dilemma: why bother investing in resources to apply metadata when the Canada site search engine was not configured to use it effectively? Yet, the search engine could not be configured and tested until there was consistent and high-quality metadata applied to web resources. Without a test environment, there was no clear proof of concept to demonstrate the return on investment of applying metadata.

4.2 Metadata Supporting Citizen-Centred Access

However, using the Canada site gateways and clusters to access GoC information and services began to resonate with Canadians. In fact, developing the site in consultation with citizens has helped to place Canada as the top-ranked country in Accenture's e-Government Leadership survey, four years in a row.[10]

Now, in order to sustain the public success of the gateway and cluster approach, the GoC has begun building a metadata content management solution which will enable the clusters to manage their links and document content by harvesting metadata directly from departmental sites. In order for this content management solution to succeed, at least a core amount of the metadata created by departments for their Internet resources must be consistent and compliant with GoC standards. So, as the Canada site becomes more familiar to users as the entry point for GoC information, it is in the best interest of the departments to improve their metadata in order to have their resources easily and properly indexed by the clusters. Using detailed controlled vocabularies across the whole GoC domain will increase access not only to the general public, but also to more specialized users as well.

4.3 Convergence

Several high-level initiatives are emerging which seek to align the structure of programs and services, as well as their information structures, across the entire GoC. The GoC has developed a Business Transformation Enablement Program[11] (BTEP) to enable business design across the government with a standards-based

https://doi.org/10.23106/dcmi.952107752

approach. The goal is to guide and expedite transformation to meet the government's high-level business objectives for interoperability. Part of the BTEP methodology is the Government Strategic Reference Model (GSRM), which allows organizations to model service delivery, including elements such as what is produced (e.g. service outputs), who produces it and who it is produced for (e.g. providers and target groups), why it is produced (e.g. target group needs, program goals and desired outcomes), how it is produced (e.g. programs, services and business processes), where and when it is produced (e.g. jurisdictions, points of service, events and cycles) and critical relationships (e.g. cross-organization value chains, accountability and performance metrics). In order to model disparate government services in this way, there will have to be an agreed-upon vocabulary for naming the components within the model. Two years ago, the Metadata Working Group created sub-groups to identify and/or create new vocabularies for use with Dublin Core metadata elements for web resources. Now, a much broader range of parties is involved in the process of determining the necessity, value and roles for creating new GoC controlled vocabularies. In fact, a broader, still-evolving understanding of the utility of the vocabularies is in progress.

Over the past year, Library and Archives Canada (LAC) has been developing a methodology called the Business Activity Structure Classification System (BASCS) for government departments to use in classifying government information or records. BASCS and GSRM use certain concepts to define government activities at various levels. For example, the GSRM has definitions for "Program", "Service", "Outcome", etc. BASCS uses constructs such as "Function", "Sub-Function" and "Activity." Recognizing this as a point of potential convergence, representatives of these and other partnering agencies have begun discussions to see how these different ways of classifying government activities might interoperate. Eventually, these activities will require interoperable controlled vocabularies to allow the specific programs to be described in "citizen-facing" terms while fitting within the government-wide model.

4.4 Building on Success

As a first step toward developing a highly integrated enterprise vocabulary, another government department, Public Works and Government Services Canada has initiated a "Metadata and Taxonomy Integration Project" which utilizes the BASCS structure and is investigating the development of a functions-based thesaurus, potentially leveraging the Core Subject Thesaurus. This will allow them to build on the concepts which exist within the CST and possibly leverage the CST expertise in ISO-based standards to build a new thesaurus.

5. Current Interoperability Mechanisms

5.1 Standards

In general, when standardized semantics and syntax are agreed upon, interoperability between systems is easier. Since the GoC has accepted the Dublin Core as its resource discovery

https://doi.org/10.23106/dcmi.952107752

metadata standard, in this context it is agreed that the form and meaning of the element "title" is defined by the DCMI. When controlled vocabularies are adopted as standard value sets, the terms within them become the recognized way of describing resources. To describe a web resource which provides "…instructions or directions (e.g. how to write a report, how to obtain a copy of a publication, how to register for a service)", those who agree to use the GoC type vocabulary would use the value "guide" in the type element, not "guideline", "handbook", "instruction", "manual", "procedure", "style guide", "toolkit", "tutorial", "user guide", etc.

5.2 Repeating Elements

The GoC vocabularies contain relatively few, high-level terms. The audience, type and format schemes are flat taxonomies. As stated in the principles, they are intended to reflect content which is found on many sites (e.g. "frequently asked questions") and/or is of significance to the GoC (e.g. audience term "seniors"). They should also allow expansion into a more detailed level of specificity, thereby complementing vocabularies created by individual departments. In fact, well before the GoC standards for metadata and controlled standards existed, federal departments were developing and using their own thesauri, schemes, vocabularies, etc. These vocabularies were built to meet the needs of the departments to organize and make accessible their own information. While they are intended to allow interoperability within a single department, or perhaps across a community within a specific discipline (e.g. Health Canada's Controlled Vocabulary[12], Canadian Immigration and Citizenship Indexing Terms[13], the question arises as to what mechanism(s) would allow them to interoperate with GoC vocabularies?

Dublin Core permits most elements to be repeated, thus allowing terms from different vocabularies to be used in separate instances of a metadata element. In the GoC, when more than one vocabulary is used, the element must be repeated and the vocabularies must be identified. Therefore, the use of terms from domain-specific controlled vocabularies along with the GoC schemes allow resources to be described more precisely while maintaining interoperability among GoC systems, such as the search engine of the Canada site and harvesting by gateway and cluster sites. This co-existence of values from different schemes to describe a single resource is currently the only strategy widely used on GoC web resources to achieve both interoperability as well as a sufficiently precise level of description.

6. Interoperability Mechanisms in Development

6.1 Namespaces

It is possible that one term could have different meanings, depending on its context. The meaning of the term "guide" in the GoC type vocabulary is not the same as a "guide" in the travel industry. The term could still be used in both contexts, as long as there is a way to avoid confusion over its meaning. It is not possible to use the same term with two different meanings in a single vocabulary, as described in section 5.1 above. However, namespaces are able

https://doi.org/10.23106/dcmi.952107752

address this issue. They provide "…the structural and semantic rules for any given data element <and> must be known if the data is to be "interpreted" correctly…" [14] When these rules are clear and explicit (i.e. publicly available), namespaces provide a mechanism to specify the form and context for the meaning of terms within a prescribed context. When a namespace is established and publicly available, they terms within them may be known and used by anyone.

6.2 Application Profiles

"An application profile is a schema, which consists of data elements drawn from one or more namespaces optimized for a particular local application."[15] (Heery: 2000). By considering each term within controlled vocabulary as a "data element", the possibility of creating a "virtual", yet still controlled vocabulary exists. Using a web service as the delivery vehicle, and maintaining the reference to the namespace from which as term is drawn, it is possible to mix and match terms from different vocabularies while maintaining their unique syntax and semantics. This is one example of the how the semantic web could be enabled via controlled vocabularies. RDF expressions of the GoC audience, type and format vocabularies have been developed in XML with the intention of allowing this type of sharing to occur. However, the profiles have not yet been placed into a GoC namespace and the web service itself is still notional.

Other ways to extend GoC schemes, such as mapping between vocabularies and/or individual terms, or developing hierarchical nested schemes are still being investigated.

7. Metadata Management Tools and Technologies

7.1 Registry

While the Controlled Vocabulary Standard offers no specific guidance on how and by whom the vocabularies should be developed, it does require them to be interoperable. It also requires that vocabularies used by the Government of Canada be registered and publicly available. Library and Archives Canada has established a registry to make standardized vocabularies available to information creators, those involved in developing and maintaining vocabularies as well as provide a centralized reference tool for use in metadata elements by GoC departments and agencies. The registry is described in a poster presented at the Dublin Core 2003 conference[16]. A list of Canadian Government-maintained Controlled Vocabularies and Thesauri is available on the Library and Archives Canada web site[17].

7.2 Training

The GoC is in the process of developing a basic controlled vocabulary training course for employees who use these vocabularies to index web resources. The goal of the course is to help them understand the characteristics and types of controlled vocabularies, the GoC metadata context in which controlled vocabularies function and give them some fundamentals and practice on content analysis. A pilot course offering is planned for late summer, 2004.

https://doi.org/10.23106/dcmi.952107752

References

[1] Speech from the Throne to Open the Second Session of the Thirty-Sixth Parliament of Canada, October 12, 1999 http://www.pco.gc.ca/default.asp?Language=F&page=informationresources&sub=sftddt&doc=sftddt1999_e.htm

[2] Common Look and Feel Metadata Standard http://www.cio-dpi.gc.ca/clf-nsi/inter/inter-06-03_e.asp

[3] Treasury Board Information Management Standard, Part 1: Government On-Line Metadata Standard http://www.cio-dpi.gc.ca/its-nit/standards/tbits39/crit391_e.asp#pur

[4] Treasury Board Information Management Standard Part 2: Controlled Vocabulary Standard http://www.cio-dpi.gc.ca/its-nit/standards/tbits39/crit392_e.asp

[5] Government of Canada Core Subject Thesaurus http://en.thesaurus.gc.ca/these/thes_e.html

[6] Final Report of the GOL Metadata Working Group<dc.type> Sub-group http://www.cio-dpi.gc.ca/im-gi/mwg-gtm/typ-typ/intro_e.asp
Final Report of the GOL Metadata Working Group <dc.audience> Sub-group http://www.cio-dpi.gc.ca/im-gi/mwg-gtm/aud-aud/docs/2003/aud-final/aud-final00_e.asp

[7] Geographic Coverage Sub-group: dc.coverage - Guidelines http://www.cio-dpi.gc.ca/im-gi/mwg-gtm/gcs-scg/docs/2002/element/element_e.asp

[8] IANA MIME Media Types http://www.iana.org/assignments/media-types/index.html

[9] <dc.format>: Guidelines http://www.cio-dpi.gc.ca/im-gi/mwg-gtm/fmt-fmt/docs/docs_e.asp

[10] eGovernment Leadership: High Performance, Maximum Value, Accenture http://www.accenture.com/xd/xd.asp?it=enweb&xd=industries\government\gove_egov_value.xml

[11] Business Transformation Enablement Program http://www.cio-dpi.gc.ca/btep-phto/index_e.asp

[12] Health Canada Controlled Vocabulary http://www.hc-sc.gc.ca/english/convocab/index.htm

[13] Canadian Immigration and Citizenship Indexing Terms http://www.cic.gc.ca/cic-index/english/index.html

[14] "Discussion Paper: XML Namespace Management within the Government of Canada" March 31, 2004, Treasury Board of Canada, Secretariat (unpublished)

[15] Application profiles: mixing and matching metadata schemas by Rachel Heery; Manjula Patel
24-Sep-2000 Ariadne Issue 25
http://www.ariadne.ac.uk /issue25/app-profiles/intro.html

[16] Encoding Scheme Registration in the Government of Canada http://www.siderean.com/dc2003/702_Poster39-registry_formatted_final.pdf

[17] Thesauri and Controlled Vocabularies http://www.collectionscanada.ca/8/4/r4-280-e.html