

Encoding Scheme Registration in the Government of Canada

Deane Henderson Zeeman, Joy Bell
Library and Archives Canada
{deane.zeeman, joy.bell}@nlc-bnc.ca

Abstract

Library and Archives Canada is the registrar of controlled vocabularies for the Government of Canada domain. In support of this role, an open source registration application has been developed.

Keywords: *Encoding schemes, Controlled vocabularies, Registries.*

1. Introduction

Library and Archives Canada (LAC) has been mandated by a Government of Canada Treasury Board Information and Technology Standard (TBITS 39.2, *Controlled Vocabulary Standard*), to be registrar of standardized vocabularies used in metadata in the federal government domain. A registry was implemented in January 2003 on the National Library of Canada Web site (www.nlc-bnc.ca). The registration form is available at www.nlc-bnc.ca/8/4/r4-300-e.php.

The function of this registry is two-fold. It provides a centralized repository of pre-determined sets of values for use as a reference tool by Canadian federal government departments and agencies for subject, audience, type or geographic coverage metadata and it makes the standardized vocabularies available to search engines and web services for validation purposes. The lists of vocabularies are available at www.nlc-bnc.ca/8/4/r4-281-e.html.

A vocabulary is registered by submitting the required information using the following rules and criteria.

2. Criteria for registering standardized vocabularies

Controlled vocabularies or thesauri may be registered. Only vocabularies developed and maintained within the Government of Canada are registered. External standardized vocabularies are not part of the registry. The DCMI registers external vocabularies.

Vocabularies must be created and maintained by trusted authorities. A *trusted authority* has a mandate within the department to develop and maintain the vocabulary. Equivalent terms must be available in both English and French. Vocabularies must be publicly available on the World Wide Web.

3. Vocabulary Titles

Standardized vocabularies are named with their official titles. Titles are provided in all languages for vocabularies that are bilingual or multilingual. Vocabularies that are derived from, modified and/or translated by someone other than the original owner are assigned a local name based on the service, project or provider name.

4. Vocabulary labels

Labels are needed for machine identification of the vocabulary. They are used as scheme names in the DC meta element.

Labels are assigned by the departmental registering agent using the following guidelines:

- labels must be unique.
- existing official acronyms or short names may be used as labels.
- official government FIP acronyms may be used in labels.
- the first two letters of the label for all schemes developed specifically for use in the Canadian federal government must be “gc”.
- no punctuation is included.

5. Who may register a standardized vocabulary?

The departmental trusted authority/maintenance agency submits a registration form to the registrar at LAC.

6. How vocabularies are registered

The vocabulary registration application makes use of LAC’s current expertise, infrastructure and the following technologies: HP UNIX TRU64 environment and utilities, open-sourced APACHE web server 1.3.27, open-sourced PHP 4.0 (Hypertext Preprocessor) programming language, data storage on ORACLE 9i database, and ORACLE utilities See Figure 1. The application works as follows:

- 6.1. A client (registration authority) completes the on-line HTML registration form at www.nlc-bnc.ca/8/4/r4-300-e.php.
- 6.2. Upon submission of the form, the data is collected, formatted and ‘~’ (tilde) delimited. An email is sent to

the Metadata Coordinator containing the formatted data.

- 6.3. The Metadata Coordinator reviews and verifies the data.
- 6.4. If the vocabulary meets the registration criteria and there are no significant errors in the registration data, the Metadata Coordinator clicks on an embedded link in the email message.
 - a. Firing the link invokes an FTP function to copy the emailed data to the UNIX file server environment.
 - b. A CRON job picks up deposited files and calls the ORACLE SQL*LOADER facility. The LOADER facility loads the data into a temporary database table in ORACLE.
 - c. The data in the temporary database table is further massaged (stripping diacritics for ease of sorting, indexing data, application of other validation) and inserted into the appropriate table structures in the vocabulary database.

- 6.5. If the vocabulary is rejected (i.e. the registration data is not understandable or the vocabulary does not meet registration criteria), the Metadata Coordinator sends an email message to the client requesting clarification or explaining the reason for rejection.
- 6.6. Uploaded vocabulary data is available immediately for client verification and public vocabulary sharing. Four lists are generated based on the DC element for which they may be used (subject, audience, type or geographic coverage).

7. Future plans

Once the DCMI Registry is publicly available, LAC will examine possibilities of interoperating with it.

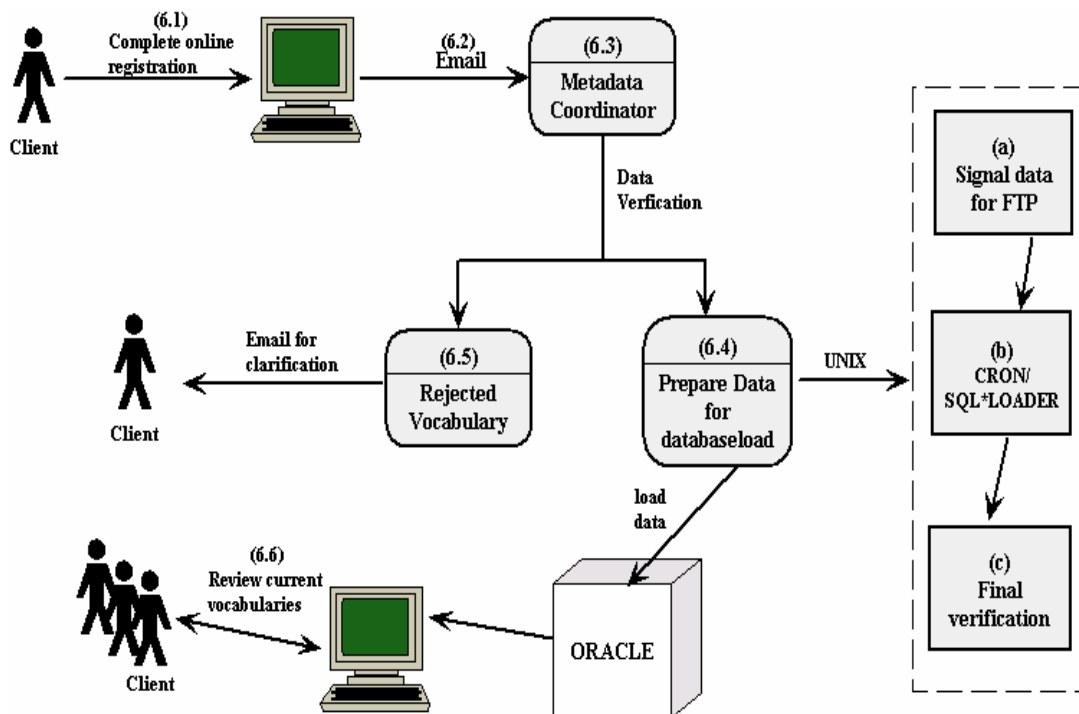


Figure 1. Registering Standardized Vocabularies in the Government of Canada