The Metadata Tagging of the Business Start-Up Assistant

Prepared by Matthew Bowler
for the Canada Business Service Centres,
National Secretariat
Industry Canada

April 19, 2004

Executive Summary

The Canada Business Service Centres (CBSC) National Secretariat undertook in 2003-04 the metatagging of all the links within the Business Start-Up Assistant (BSA), a cluster within the Business Gateway.

Using a metadata tool built into the BSA's link management system, the project's two metadata "taggers" applied or verified ten metadata elements according to the Business Start-up Assistant Metadata Guidelines for the CMS Project. The taggers also performed a number of content management and quality assurance tasks. The taggers achieved a daily tagging average of 24.3 link records per person (or 18½ minutes/record), with sustained periods where averages in excess of 37 records per day (12 minutes/record) were achieved. These figures include most non-metadata tasks performed by the taggers, such as fixing broken links, correcting spelling errors in the database, etc., but exclude management and supervisory activities performed by other members of the BSA Team. A link record includes both the English and French versions of the link.

These results can be generalized to make rough projections for other metadata projects, provided that several particulars and variables are taken into account, including but not limited to the number of metadata elements to be tagged, the capabilities of both the taggers and the metadata tool they employ, and the nature of the group of links to be tagged.

1. Objective of the Report

The objective of this report is to document the processes, resources and tools used for the application of metadata to the links in the BSA, but moreover, to measure metadata tagging activities. The report provides a benchmark for similar projects and highlights the numerous factors that must be taken into consideration when measuring and evaluating tagging productivity.

2. The Project

The Business Start-Up Assistant is a flagship product of the Canada Business Service Centres and one of the most popular clusters on the Business Gateway site. The BSA combines start-up information from the federal, provincial and territorial governments, the private sector and other sources. Information in the form of links is organized by topics such as market research, financing, taxation, business planning, and many others.

In this project, which spanned 87 workdays, two taggers applied metadata to over 6000 unique links contained in approximately 3600 link records.[1] BSA links are organized under 33 topics for each of Canada's 13 provinces and territories, totalling 429 web pages. The BSA contains a diverse assemblage of links but does contain patterns, such as the presence of numerous documents from the CBSC Business Information System (BIS) database[2] and the organization of links by topic and province. These characteristics can be leveraged to increase tagging output by using a link management/metadata tool with bulk metadata application capabilities.

Apart from links, the BSA content pages were also tagged in this project. However the requirements for tagging these pages and the involvement of taggers were different from those for

[1] Web resources often have both French and English versions; each version has its own URL, but both are housed in the same link record in the database. All discussion of tagging output in this report is calculated based on link *records* as opposed to unique links.

[2] A version of each BIS document generally appears for each province/territory, although they are basically identical. Using bulk functions these documents can be tagged quickly in groups of thirteen.

the links on pages. Therefore, they are excluded from the daily tagging average. A brief description of the metatagging of the BSA's content pages is provided in Section 7: Results.

3. The Metadata Team

The two metadata taggers on this project were contractors working as part of the BSA Team. They had regular contact with the BSA Content Manager, who was responsible for the day-to-day management of the project. The BSA Cluster Manager was responsible for the preparation of guidelines, training and the overall direction of the project.

Skills

Both taggers had previous Web content management experience, and one also had some metadata experience. Both were familiar with the link management tool, although neither had previous experience with the BSA.

Training

The taggers attended the two-day Canada Metadata Forum in September 2003 and received in-depth training from the Content Manager. Training provided the taggers with an understanding and appreciation of the concepts and principles of metadata as well as the unique requirements of the BSA. The BSA metadata requirements were documented in the Business Start-up Assistant Metadata Guidelines for the CMS Project. These guidelines were updated when additional rules were required to handle situations not initially covered by the BSA guidelines or explained the Metadata Implementation Guide for Clusters and Gateways.

Language

French and English metadata was applied for all applicable fields, regardless of whether the source document was bilingual or unilingual English or French. One tagger was an English native speaker, one was a French native speaker. Both were capable of researching the vast majority of pages in their respective second languages.

The language requirements for the various metadata and other fields varied, which was reflected in the different input methods for these fields (pick lists and text boxes). For pick list fields, pre-defined translations from controlled vocabularies were automatically provided by the link management tool. Fields such as dc.title, author and more importantly dc.creator and gcms.caption, were entered using text boxes, which required that the taggers enter the information in both languages separately. See Table I below for an overview of language requirements for the different fields.

In the case of gcms.caption, the division of work was asymmetric. The French native speaker wrote his own captions in both French and English. The English native speaker wrote his own captions in English and proofread the English captions of the French native speaker, but left his own French captions for the French native speaker to write.

It should be noted that both the link management tool and the metadata team were well equipped to meet the language requirements of this project. A project employing a unilingual team and/or a link management tool without bilingual metadata functionality will encounter serious obstacles. For example, the use of a translation service would increase the cost of the project, cause processing delays and would require the addition of a verification step to ensure that the translation was appropriate for the context of the link and/or the BSA topic.

4. The Link Management/Metadata Tool

The taggers made use of the Portal Publishing, Links and Content Management Solution by Alika Internet Technologies,[3] developed on a Lotus Notes platform and incorporating full metadata functionality.

This link management tool was designed to meet the needs associated with the development, implementation, update and maintenance of "portal" Web sites. Each link occupies a single link record in the database, although one link can be displayed in up to 60 different locations on the Web site by applying different heading structures to the same unique link.. Link documents and content pages can be added or modified, then published to the Web site quickly.

The metadata tool is fully integrated within the link management tool, and includes the following features for applying metadata:

- Pick lists for faster and more accurate input of data.
- Metadata can be assigned to links in bulk based on URL patterns or other criteria.
- Link Checker function can harvest metadata from the source Web sites.
- Metadata can be input into the database through sharing with other clusters.

Development

Various improvements to the link management tool's metadata functionality were implemented over the course of the project, meaning that some important features (bulk search and replace, auto-dc.creator) were not available at the outset. The implementation of such improvements resulted in short-term setbacks, but ultimately improved the taggers' productivity.

Sharing/Harvesting

The BSA, together with four other Business Gateway clusters, participates in a Central Metadata Repository created by Alika Internet Technologies. This relationship allows the BSA's link management tool, to import *shared* metadata created by the other clusters, thereby reducing duplication and saving time. The tool is also able to *harvest* metadata directly from the link resources. Shared and harvested metadata values had to be verified before they were accepted as final values.

For many BSA links no shared or harvested metadata were available. For example, harvested dc.subject values were available for 27% of BSA links, while shared dc.subject values were available for 14.5% of the links. Harvested dc.description values (which are easily adapted for use as gcms.captions) existed in at least one language for 34% of BSA links; shared English dc.descriptions existed for about 7% of BSA links. Except in the case of descriptions, harvested metadata was only occasionally appropriate for use. While rarer, shared metadata was much more reliable, and presumably as more Business Gateway clusters tag more links, its usefulness will improve.

5. Metadata

Taggers were responsible for tagging or verifying the five mandatory Dublin Core metadata elements[4] for links as specified for Government of Canada clusters and gateways: dc.title, dc.creator, dc.language, dc.subject and dc.identifier (the link's URL). The taggers also applied the following metadata elements: dc.audience, dc.type, dc.coverage.spatial, gcms.caption (an abbreviated description), and gcms.creator.type. The element dc.date was automatically harvested

[3] The descriptions in this section are adapted from Alika's Web site. For further information see http://www.alika.ca/portal_publishing_applications_e.html.

[4] As outlined in the CMS Metadata Working Group's Metadata Implementation Guide for Clusters and Gateways, April 2003.

from the source documents if available but otherwise was not applied. Table I, below, provides an outline of the metadata elements and their tagging requirements.

Values for dc.identifier (URL), dc.title and dc.language were already present in the database records and only needed to be verified and occasionally modified. The elements dc.coverage.spatial and dc.audience were tagged when appropriate; the dc.audience value "business" was applied to all links by default and other dc.audience values were added when applicable. The remaining metadata elements were applied to all links. Controlled vocabulary values for dc.subject, dc.type, dc.audience, dc.coverage.spatial and gcms.creator.type were entered using pick lists while text elements (dc.title, author, dc.creator and caption) were entered using text boxes.

The metadata elements that took longest to tag were dc.subject, gcms.caption and dc.creator. For dc.subject the taggers applied a maximum of eight dc.subject terms or values drawn from the following prioritized list of vocabularies:

1. Terms extracted from the Government of Canada Core Subject Thesaurus (CST) that matched the names of BSA topics and sub-topics
2. BSA-specific controlled vocabulary containing over 80 terms extracted from the CST
3. CBSC-specific controlled vocabulary containing hundreds of terms also extracted from the CST

The taggers had to read and understand each document, then visually scan the pick lists for appropriate dc.subject values (with experience this scanning could be done mentally).

The element gcms.caption was often excerpted from text in the document itself, or from shared or harvested metadata descriptions gathered by the link management tool. When no excerpted caption or shared/harvested description was available, the taggers wrote their own caption. There were unique language issues in the gcms.caption field, which are discussed above in Section 3: The Metadata Team, "Language."

For dc.creator, shared values or values from the auto-dc.creator database (see below) were often used, which required very little time. However, when shared or auto-dc.creator values were not available and taggers had to research and enter appropriate values in both official languages, dc.creator could be time-consuming.

Table I

Metadata and Other Elements					
Element	**Input Method**	**Tagged or Verified**	**Time***	**Second Language**	**Sharing & Harvesting**
Dublin Core Elements					
dc.title (EN & FR)	text	verified	1	manual	-
dc.audience	pick list	sometimes tagged	1	auto	yes
dc.subject	pick list	always tagged	4	auto	yes
dc.type	pick list	always tagged	2	auto	yes
dc.creator	text	always tagged	3 (1)	manual	yes & auto
dc.coverage.spatial	pick list	sometimes tagged	1	auto	yes
dc.language	check box	verified	1	-	-
dc.identifier (URL - EN & FR)	text	verified	2	manual	-
dc.date	auto	automatic	0	-	harvested if available
GCMS Elements					
gcms.caption	text	always tagged	4	manual	descriptions
gcms.creator.type	pick list	always tagged	2	auto	sometimes auto
gcms.topic	auto	automatic	0	auto	-
Miscellaneous Elements and Fields					
indentation (UL)	check box	verified	3	auto	-
PDF disclaimer and link	check box	verified	2	auto	-
heading structure	variable	verified	3	auto	-

*Values in the "Time" column provide a rough estimate of the time required to tag the corresponding element. A value of "1" would be under 10-15 seconds, while "4" could sometimes mean over 10 minutes.

Automatic dc.creator

The auto-dc.creator function was integrated into the link management tool during the project's first month. Created for the same group of Business Gateway clusters mentioned above, this function is an extension of the "Metadata Lookups Database," which contains pick list controlled vocabularies and standardized domain names. URL domain names for all BSA links were automatically checked against an authoritative list of domain names; when a match was found the corresponding dc.creator value would appear in an auto-dc.creator window. These values could be selected in bulk as default final dc.creator values, or could be accepted on a case-by-case basis. Over 75% of BSA links had corresponding auto-dc.creator values, and while some values required modification, the BSA team did make regular use of auto-dc.creator values. Also built on the Metadata Lookups Database, an auto-gcms.creator.type[5] function was being implemented; at last count 14.5% of BSA links had auto-gcms.creator.type values, largely confined to federal and provincial government links.

Cluster Administrative Elements and Quality Assurance Tasks

To capitalize on the fact that applying metadata required that the taggers open and examine each link record individually, they were also responsible for performing quality assurance on a variety of content and record management issues, such as identifying and correcting dead links, duplicates,

[5] Some examples of gcms.creator.type values are "Federal Government," "Association" and "Non profit."

improper link placement, redirected URLs, and title changes. Some problems required significant time spent researching or consulting with the Content Manager. It should be added that from the taggers' perspective, non-metadata tasks added welcomed variety to the project.

6. The Approach

The BSA Content Manager assigned unique blocks of 33 topics/sub-topics to each tagger. This decreased the possibilities of the same link record being accessed at the same time which would result in database conflict problems. The taggers worked through each province for each topic, one record at a time. When the search and replace function became available, it became possible to bulk tag a variety of metadata elements prior to opening the first record for a topic/province section. Also, this function allowed for the bulk application of metadata for CBSC BIS documents – generally a version of each BIS document exists for every province/territory. For this type of link, taggers could tag 13 records in the time it would normally take to tag 2 or 3 records.

Alternate approach

In retrospect, had the bulk application function been available from the outset, it would probably have been faster to spend a few weeks simply doing as much bulk tagging as possible before beginning the topic/province linear approach. For example, using the link management tool's search function the tagger can bring up all records containing ".gov.pe" (which yields 44 hits in the BSA database), tag these records with "dc.coverage.spatial=Prince Edward Island" and "gcms.creator.type=provincial-territorial government."

There are numerous ways one could run searches (using only URL and title), then bulk apply metadata based on the search results. After the bulk tagging opportunities are exhausted, the tagger would then begin checking each link individually and filling in any fields which had not been bulk tagged. Such an approach could be used to apply a great deal of metadata - every metadata element except gcms.caption/dc.description could be at least partially applied in this way, although some links would not turn up in any searches and would receive no bulk tags.

There are, however, some problems with this approach. With more than one person working in the database, and using the link management tool's bulk functions, occasional database save conflicts would be likely. Also, applying metadata in bulk entails spending time waiting for the link management tool to process the bulk application. Finally, it must be emphasized that this approach is intended to speed up the application of metadata, but does not obviate the need to closely examine the content of each document. Metadata applied in bulk and without real knowledge of the source document should be considered tentative until the document has been examined. Nonetheless, an approach which incorporates judicious bulk tagging should offer a substantial advantage over the linear approach.

7. Results

When days on which no tagging was done are excluded, the average daily metadata tagging output for this project was 24.3 records per person, or 18½ minutes per record. When all days are included the overall daily average is reduced to 21.5 records per person. The taggers' individual outputs differed only slightly, and this difference can be explained by the linguistic division of labour for gcms.caption (see Section 2: The Metadata Team "Language," above).

The overall average of 21.5 records per person/day includes non-tagging days, when taggers were occupied with activities such as training, discussion with the Content Manager, meetings, revisions, etc. Because this report only examines work performed by taggers; any costing analysis

of metadata tagging would also have to take into consideration management and supervisory resources applied to this task.

Appendix I displays the taggers' output over 87 days. The data upon which the chart is based contains daily averages ranging from the single digits up to 87 records for one tagger. Over the short term, this variability is due to tagging groups of BIS documents, gcms.caption translation tasks, and non-metadata tasks. In order to reduce this variability, each of the chart's vertical bars averages the two taggers' outputs over three days.

When the short-term variability is removed through the three-day, two-tagger blending and averaging, four prominent features which can be connected to events in the course of the project become apparent:

- The initial learning curve is visible between Days 1 and 11. During this period the taggers were building their knowledge of the project's metadata elements and controlled vocabularies, as well as the BSA's organization and subject matter. Output averaged only 9.4 records per day during this period, although the output for Days 12 to 15 exceeded the project's overall average, indicating that the learning curve had been largely overcome by the project's third week. At approximately the same time the metadata tool's bulk search and replace functions became available, raising the output further.
- Major revisions of previously tagged gcms.caption and dc.creator values began around Day 21, resulting in a significant decline in metadata output for several days. This was partly due to the introduction of the metadata tool's auto-dc.creator function. While it caused a short-term setback, the auto-dc.creator function increased productivity once it was implemented.
- A peak output period, during which no major disruptions occurred, stretches from approximately Day 30 to Day 60.
- At some time around Day 61 a major review of content-related changes was initiated, which resulted in another significant decline in metadata output.

In order to arrive at a useful figure for sustainable peak metadata output, erratic events such as the initial learning curve, changes in the link management tool, revisions, and major content management tasks must be factored out. Clearly, any project will experience unforeseeable disruptions, but strictly speaking the larger disruptions in this project did not pertain to the application of metadata, and as such can be excluded from this calculation of sustainable metadata output.

The highest 10-day average (recall that these figures already aggregate the outputs of two taggers) is 37.4 records per day (or 12 minutes/record), this can be viewed as an optimistic average, while the highest average over 20 days of 35.1 (12¾ minutes/record) can serve as a conservative figure.

Tagging BSA Content Pages

The BSA's content pages, numbering almost 500, were also tagged in the course of this project. For technical reasons, and because some of this work was performed by members of the BSA Team other than the taggers, the quantitative recording of this aspect of the project is not accurate enough to allow a detailed analysis. Also, since the content pages were tagged after the sample period used for links, this phase of the BSA metadata project is entirely excluded from the analysis in this report. Nonetheless, the available data suggest that content pages required much less time to tag than links.

Apart from the use of dc.description instead of gcms.caption, the same metadata elements were tagged for content pages as had been tagged for links. Discretion was required in the creation or

selection of dc.description and dc.subject values for content pages, but other metadata values were either identical across all pages or followed predictable patterns. Since each topic page is repeated across all 13 provinces/territories, only 33 different descriptions and sets of dc.subject values had to be assembled, although the taggers had to insert province/territory names into many of the descriptions. Tagging content pages was a task that lent itself extremely well to bulk tagging.

8. Conclusions

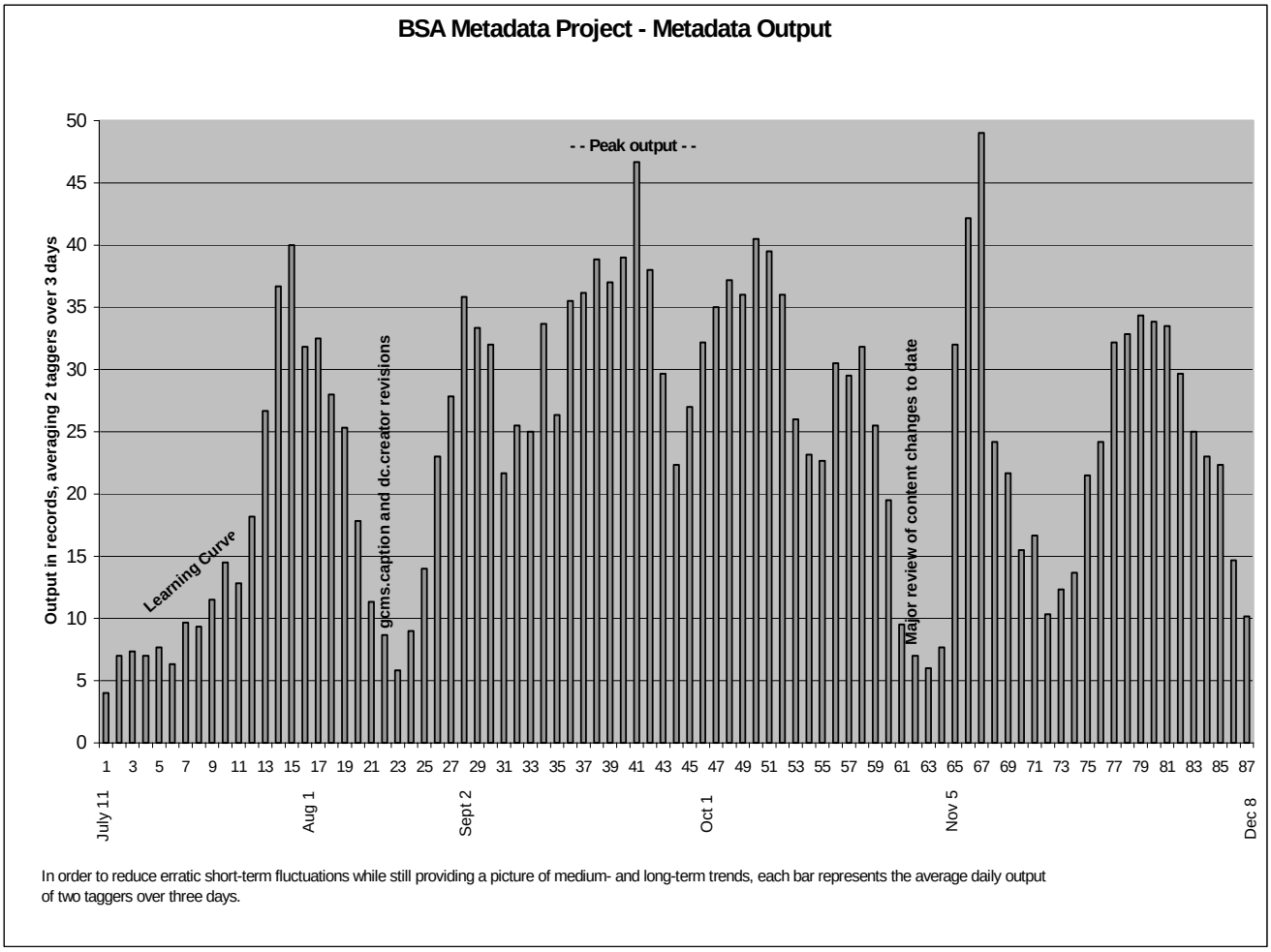
This project entailed applying a variety of metadata elements to a group of 6000 fairly disparate external links contained in the BSA database, and required that the taggers perform some non-metadata quality assurance duties. The taggers' task was simplified considerably by a link management tool with substantial metadata functionality. The BSA metadata project can be used as a benchmark for other similar clusters, but the current project's daily output of 24.3 records per person should be used with caution. The evaluation of any metadata project will need to take into account a number of variables which will have an impact on tagging productivity.

Some of these variables include:

- the number of metadata elements to be tagged.
- the capabilities of the metadata tool employed – specifically, integration with the record management system, pick lists and other input functions, bulk application, and sharing/harvesting capabilities.
- whether basic metadata elements such as dc.identifier (URL), dc.title and language are prepopulated in the database, or must be entered.
- detailed cluster-specific guidelines for the application of metadata to elements required for the cluster
- taggers' familiarity with the record management/metadata tool.
- taggers' familiarity with the subject matter and the group of records to be tagged.
- taggers' prior understanding of metadata concepts and controlled vocabularies.
- the level of participation of supervisors and management.
- whether taggers are responsible for any content management or quality assurance tasks to be carried out at the same time as the tagging.
- whether the project involves tagging link records or content pages – there are various reasons why links will take longer to tag than content pages, i.e. dc.creator will be identical across all content pages in a cluster, whereas taggers must research a wide variety of dc.creator values for links.
- the organization of the group of records to be tagged, particularly the existence of inherent and identifiable patterns which will facilitate bulk application (assuming the project uses a link management tool with bulk application capabilities).
- language requirements, and the extent to which the project team and the link management tool are equipped to meet these requirements.

This report on metatagging metrics documents the practices and outcomes of one metadata project, provides a rough benchmark for other projects, and endeavours to highlight important issues. While metadata quality should always be a key priority, productivity and efficient processes are of special importance given the vast number of links in clusters and gateways and the labour-intensive nature of metatagging.

Appendix I



In order to reduce erratic short-term fluctuations while still providing a picture of medium- and long-term trends, each bar represents the average daily output of two taggers over three days.