GOL Clusters and Gateways Content Management Solution Proof of Concept Taxonomy Project

Presentation to Knowledge Taxonomies

GSLIS, McGill University

November 3, 2003

By Elsa Van Hulst, 2003 GSLIS graduate

Presentation Overview

- **Context of Government On-line Clusters and Gateways**
- **Implementing a common Content Management Solution (CMS)**
- **Preparing for the CMS – shared processes and standards**
- **Need for a common organizational structure to describe the collective content**
- **Proof of Concept Taxonomy**
    - Business objective
    - Benefits
    - Taxonomies vs thesauri – Controlled vocabularies
    - Results
    - Challenges

Context: GOL Gateways & Clusters

Canada Site

A non-departmental approach to finding government information.

The 35 clusters are organized into primary topics of public interest or targeted audiences.

Context: GOL Cluster Goals

- **Clusters are about Improving Access**
  - One-stop shopping, no-wrong-door to information, online filing, transactions, consultations etc.

- **Their approach is to build horizontal relationships and deliver client-focussed information and services**

- **Moving towards integrated information and service delivery across jurisdictions**

Canada

A Common Content Management Solution (CMS)

- **Move towards shared technology, shared processes, shared standards…**
  - Increased interoperability, increased efficiency
  - … service integration

Metadata is key to service integration… metadata is about making connections…

Working Together

- **Group dynamics**
- **Defining the business processes and agreeing upon the metadata standards**
- **Strong personalities, highly dedicated people that have put a lot of efforts into their cluster projects and into the CMS**
- **Commitment**

Metadata – Shared Semantic Model

- **Based on the Dublin Core metadata standard**
  - Dublin Core defines the semantics (meaning) of 16 elements
  - A controlled element : the content or value is
    - formatted in a standard way
    - a term from a specified list of acceptable values or controlled vocabulary
  - Controlled vocabularies should be named, registered and identified in metadata records as schemes
- **Extended with other standards in order to satisfy specific functions and domains**

Government of Canada Metadata Framework

Record keeping metadata requirements, based on GILS and Dublin Core developed for GoC.

Domain specific metadata sets extending the Dublin Core exist or may be developed.
Specific international domain specific schemes are used in GoC e.g. geospatial (ISO 19115), learning materials (IMS, SCORM.), audiovisual (MPEG7).

Domain specific metadata

Extending the Dublin Core to provide metadata guidance for management of clusters and gateways.

Dublin Core
CLF

Record Keeping

Portal Administration

Common Look and Feel Standards: five metadata elements for GoC web-sites

Dublin Core adopted as GoC Standard

Dublin Core:
Metadata Standard for Resource Discovery

Title * *(dc.title)*

Creator * *(dc.creator)*

Language * *(dc.language)*

Date * *(dc.date)*

Subject * **(dc.subject)**

Description (dc.description)

Publisher (dc.publisher)

Contributor (dc.contributor)

*CLF (5 elements)

Type (dc.type)

Format (dc.format)

Identifier (dc.identifier)

Source (dc.source)

Relation (dc.relation)

Coverage (dc.coverage)

Rights (dc.rights)

Audience (dc.audience)

Dublin Core

CLF

16 elements

- **Only the five elements previously approved as Common Look and Feel (CLF) Standard 6.3 are mandatory across all GoC Web sites**
- **The Dublin Core is used when more than 5 elements are needed.**
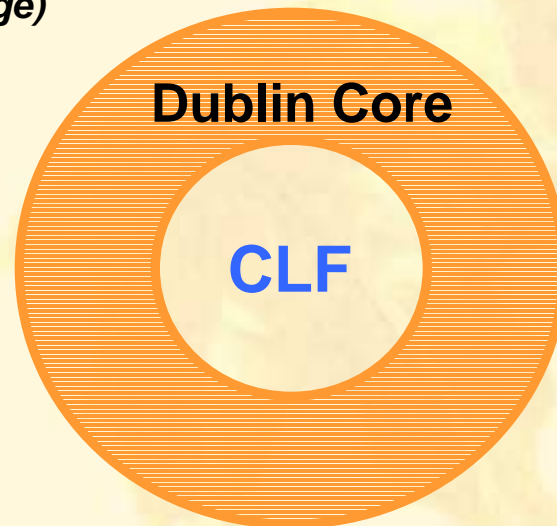
Evolving Principles for Developing GoC Schemes

- **Applicable:**
  - terms represent content found on a significant number of Government of Canada Web sites, and/or are of substantial significance to Government of Canada programs/services
- **Recognizable:**
  - terms are understandable by implementers/indexers
- **Unique:**
  - no terms will be a synonym of an existing term [within one scheme]

Usage of Schemes to Support Access

- **Metadata elements can be considered "facets" for organizing government Web sites**
  - Audience
  - Geographic Coverage
  - Date
  - Type

  and
  - Subject

Audience

Type

Combined facets: Type and Audience

National Library of Canada / Bibliothèque nationale du Canada — nlc-bnc.ca

Canada

Français | Contact Us | Help | NLC Search | Canada Site

NLC Services:

Directory of Services

Services for Educators

Services for Researchers

Services for Libraries

Services for Publishers

Discussion Lists

Services

Services for Educators

For Teachers (Parents Welcome!)

Services for Researchers

Guide to Services for Researchers
Reference Service
Canadian Children's Literature Service
Canadian Literature Research Service
Canadian Newspapers on Microform Held by the NLC
Canadian Theses Service
Genealogy and Family History
Jacob M. Lowy Collection
Music Division
The Rare Books Collection

Schemes for Interoperability Within GoC

- **Accommodate diversity**
    - Broad, high-level GoC-wide schemes
    - More specific domain or departmental schemes

Wide-spread use of common schemes is key to interoperability

Standarizing as Much as Possible but Allowing for Cluster-Specific Metadata

Mandatory CLF5

Dublin Core:

dc.title

dc.creator

dc.subject

dc.date

dc.language

Plus

Gateway and Cluster Metadata

Additional Dublin Core:	Cluster Contextual:
dc.identifier	gcms.cluster
dc.description	gcms.topic
dc.coverage.spatial	gmcs.caption
dc.type	gcms.creator.type
dc.audience	gcms.contact.public
dc.format	
dc.contributor	
dc.publisher	
dc.relation	
dc.rights	

gmcs.cluster, gcms.topic, etc. —

Cluster Specific Metadata Elements offers Flexibility and Adaptibility, However…

Cluster-Specific Subject Classification

- **Cluster mappings do not currently produce equivalent results**
- **An approximation of "no wrong door"**
- **Indicative of possible sharing opportunities across clusters and reduced duplication**
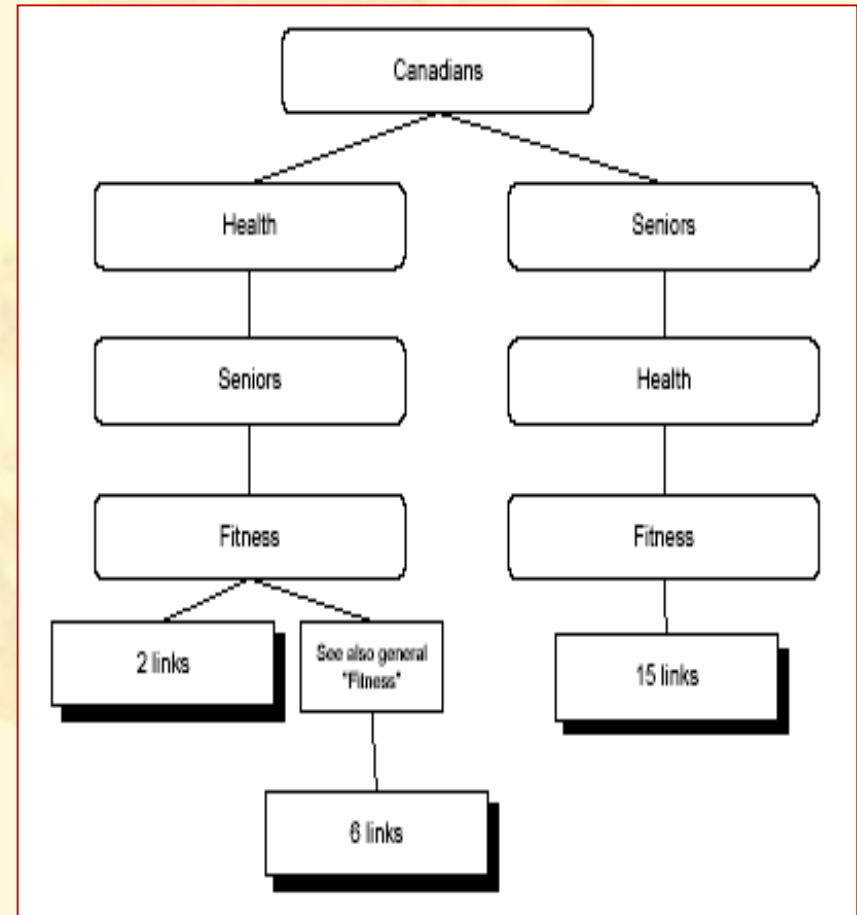
Proof of Concept Taxonomy – Objectives

- **Define cluster topic connections and create a classification for organizing cross-cluster content**
  - share efforts related to resource discovery and partnership management
  - reuse metadata
  - enable dynamic delivery of potential cluster content (based on CMS inventory which will include many harvested links and associated metadata)

Benefits of Having a Common Taxonomy

- **More effective information retrieval and better navigation**
  - Ability to refine all-of-government searches against hierarchical subject structures
  - Move towards a consistent view to and organisation of information portrays external credibility and leadership
- **Discovery of content connections, affinities, gaps in content offerings, opportunities, …**
  - Accelerate the connections of people through content triggers
  - Reduce duplicate work
  - Creates organisational synergies, which reduce costs and increase innovation across organisational boundaries

Evolving Use Case for Taxonomies?

…related content, variant views to a topic, related initiatives, people trying to achieve similar goals, etc

Discovery of connections

…the potential for discovery of connections may extend beyond just Clusters and Gateways

Information retrieval

Information delivery

…vision of horizontal management within the Government of Canada…

…Evolving Taxonomy Strategy

- **a flexible front-end taxonomy (cluster site navigation),**
  - supports client-centered information retrieval
- **operating on a stable, controlled vocabulary infrastructure**
  - supports the need for making connections within the organization
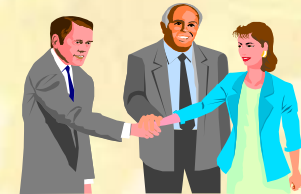
Organizing Information to Meet User Needs

Citizen-centric

Fluid taxonomy

Network of thesauri

Search

Metadata creation

Metadata
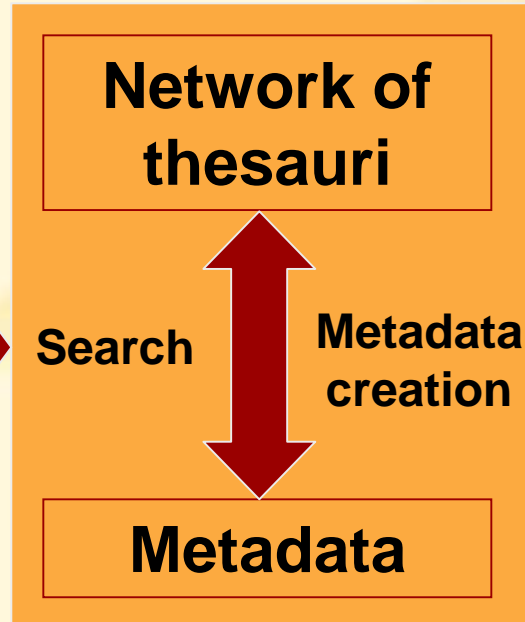
Stable controlled vocabulary

By recognizing differences in language

Evolving Use Cases, but Down to Old (Somewhat Unpopular) Methods?

- **Controlled vocabularies: Taxonomies, Thesauri**
- **So what about dc.subject?**
  - Must contain terms from a controlled subject vocabulary
  - At least one of which coming from the Core Subject Thesaurus
- **Government of Canada Core Subject Thesaurus**
  - Limited and unequal coverage
  - Mostly high-level
  - Few relationships between terms, not browsable
- **Other thesauri in use – Health, Statscan, …**

Nesting and Merging Multiple Thesauri

- **Done in many private corporations, especially when companies merge**
- **Glaxo Wellcome (now GlaxoSmithKline) case; merging of various thesauri into a sort of mega-thesaurus providing unified access to many different repositories***
  - Preferred terms (53,000?)
  - Lead-in terms (201,000?)
  - Hierachical or associative assertions (443,000)

More than just a thesaurus by virtue of all those defined relationships?

Top-down vs Bottom-up

- **Metadata (foreseeable information needs)**
  - Manual tagging (pre-determined facets and terms)
  - Automatic tagging by auto-categorisation software (not necessarily based on pre-determined facets and terms)
    - Statistical clustering
    - Rule-based
- **Data mining**
  - Knowledge discovery for areas where you have less clear a sense of what you're looking for
  - Low cardinality (too many similar values associated to limited number of facets), transaction data, DNA sequences

Proof of Concept Taxonomy: Results of our Efforts

- **A structure that might serve as a basis from which we can start building a more comprehensive and hierarchically-structured controlled subject vocabulary**

- **Clusters learn about connections with other clusters during the process**

- **Clusters discuss classification principles and terminology issues which can lead to more consistency across cluster navigation taxonomies**

Taxonomy Challenges

Current challenges:

- **Determining what do we want to provide to our users in terms of search interface**
- **Political tensions**
- **Managing expectations**

Foreseeable challenge if the Proof of Concept goes any further:

- **Convincing those in power to invest in a controlled subject vocabulary, not just IT**

Allison gets a customized information and services stream (syndication)

Allison's agent defines delivery priority for info streams

Allison receives government services as an integrated, cross-jurisdictional package

Inter-operability of infrastructure and technologies

Transformed, integrated and efficient service delivery

Shared standards and business processes

Government of Canada

Gouvernement du Canada

Canada