Creating and Managing Controlled Vocabularies for Use in Metadata

Tutorial 4
DC2004, Shanghai Library
14 October 2004

Stuart A. Sutton & Joseph T. Tennis
Information School of the University of Washington, Seattle, U.S.A.
Goals of the Tutorial

• Controlled vocabulary development [Joe Tennis]
  – Development considerations
  – Development techniques

• Managing controlled vocabularies for the Web [Stuart Sutton]
  – Vocabulary and vocabulary term identification
  – Vocabulary declaration
  – Vocabulary publication
Contents

• **Definition:** What do we mean when we say “controlled vocabulary”
• **Issues:** What are the issues in developing, generating, using and managing vocabularies?
• **Creation:** How are controlled vocabularies created?
  – Creating a new controlled vocabulary
  – Creating a useful controlled vocabulary for a community already using many different vocabularies
• **Form and Use:** How are controlled vocabularies represented and used?
  – URI references for identifying vocabulary terms
  – Human-readable strings that represent vocabulary concepts
• **Management:** How are controlled vocabularies managed?
  – Vocabulary declarations and publication
  – Simple and complex registries
Issues

• No Controlled vocabularies are used. Many metadata initiatives do not use controlled vocabularies in generating metadata values

• Vocabularies used are not identified. Even when controlled vocabularies are used, they may not be identified in instance metadata

• Vocabularies are not publicly accessible. Even where the vocabulary is identified in instance metadata, public access to the vocabulary by humans and/or machines is unavailable

• Vocabulary identification is lost in “dumbing down”. Even when identified in complex instance metadata (e.g., qualified Dublin Core), the identifications are lost in the process of “dumbing down”
Definition

• Controlled vocabulary (CV): A finite set of distinct values for a metadata property
  – Different from a “metadata vocabulary” which defines a finite set of properties (i.e., a schema)
  – In the metadata statement “dc:subject=cybernetics”, we are concerned only with the controls placed on the right-hand side of the statement (i.e., the scheme or value space)
• Not just for the “subject” property!
  – With the exception of properties with uncontrolled value strings in Dublin Core (e.g., dc:description), all properties can successfully use controlled vocabularies to increase precision and enhance meaning
  – E.g., the DCMI vocabulary for use with the dc:type vocabulary is a “controlled vocabulary”
Controlled Vocabulary Development
Vocabulary Development

• Two types of development
  – Creating a new controlled vocabulary
  – Creating a useful controlled vocabulary for a community already using many different vocabularies
Developing Vocabularies

• Creating a new controlled vocabulary
  – Gather data from your community combination of ways
    • Representatives speak for users (vet ideas in committee)
    • Gather information about the community and its information needs
    • Gather user search data
  – Construct vocabulary
    • Identify terms
    • Identify relationships between terms
  – Maintain vocabulary
    • Keep it current and useful!
    • Based on user, domain, and representative input
Creating a New Controlled Vocabulary

• Gather data from your community [1/3]
  – Representatives speak for users
    • Committee meetings
    • Draft vocabularies
    • Vet vocabularies to group
  – Good things about this approach
    • Faster
    • Cheaper
  – Cons of this approach
    • No real user data
    • No way of knowing whether what you built works
Creating a New Controlled Vocabulary

• Gather data from your community [2/3]
  – Information needs of a community
    • What terms are in the documents they use?
      – E.g. web pages, journal articles, maps?
    • How is the community structured?
      – Who creates information, who consumes it, who stores it?
      – Are there standards (educational standards for example) that shape information flow?
      – Are there basic philosophical assumptions about the information flow? (does it all need to be captured or just part of it?)
Creating a New Controlled Vocabulary

• Gather data from you community [3/3]
  – What search terms do they use?
  – What tasks do they perform that they need information for?
  – What habits do they have for searching for information (other people?) - how does that affect how you build your vocabulary?

  – All three of these data collection activities will help you construct the terms and relationships between terms.
Creating a New Controlled Vocabulary

• Construct Vocabulary
  – Identify terms
    • Pull together synonyms, disambiguate homographs
    • Any term that is not “official” can be used as an aid in search (expanding the query to direct it to the “official” term
  – Identify relationships between terms
    • What are the relationships between terms that will aid the user during search and retrieval?
    • Make those relationships explicit in your metadata.
  – These actions make a controlled vocabulary “controlled”
Creating a New Controlled Vocabulary

• Maintain the vocabulary
  – Check search logs…see where search is successful and where it fails
  – Where it fails, adjust the vocabulary
  – Add new terms as needed
  – Take out terms as they are not needed
  – Think about using those terms as query expansion
Developing Vocabularies

• Creating a useful controlled vocabulary for a community already using many different vocabularies
  – Merging vocabularies
  – Create high-level “switching” vocabulary
  – Create a third new vocabulary and link to similar terms in the existing vocabularies
Developing Vocabularies

• None of these options are ideal

• All three options require careful understanding of the community (domain) and the use/users of the vocabularies

• Addressing the terms in these vocabularies with URIs helps disambiguate the individual terms, because these terms will be defined in namespaces
Developing Vocabularies

• Merging Vocabularies
  – The vocabularies (terms and relationships between terms) can be merged in order to facilitate search
  – The structure of each vocabulary must be accounted for in the merging
  – We must account not just for the occurrence of similar or the same terms, but also the meaning of those terms as they’re used by the indexers and the searchers
Developing Vocabularies

• Create a high-level “switching” vocabulary
  – This ignores most of the relationship structure of the existing vocabularies
  – Imposes a broad “umbrella-like” structure to the vocabularies
  – Easier to do (with an accurate understanding of the vocabulary, domain, and users), but search may suffer a bit (potential loss of search control)
Developing Vocabularies

• Create a third new controlled vocabulary and link its terms to similar terms in existing vocabularies
  – More time intensive and costly than the other two options, and more involved than crafting a vocabulary from scratch
  – It is costly because you have to monitor all the changes made to each of the linked vocabularies, and you must add these to the search query expansion in a thoughtful way, requiring a lot of time and effort
Infrastructure to Support Exposure and Use of Vocabularies

Identifying, Declaring and Publishing Vocabularies
Webized Controlled Vocabularies

“Webized” controlled vocabularies \textit{and} vocabulary terms are:

– Persistently and uniquely \textbf{identified} URIs for names/tokens/identifiers

– Formally \textbf{declared} by means of a schema language
  Represented in XML or RDF/XML

– Made Web-available by being \textbf{published}
  Published through a Web-accessible registry
Vocabulary Term Identification

• **Assertion 1:** All webized controlled vocabulary terms must be assigned a URI reference

  “The most fundamental specification of Web architecture … is that of the Universal Resource Identifier, or URI. The principle that anything, absolutely anything, ‘on the Web’ should be identified distinctly by an otherwise opaque string of characters … is core to the universality [of the Web].”

  Tim Berners-Lee, “Web Architecture from 50,000 Feet”
  <http://www.w3.org/DesignIssues/Architecture.html>

– Required by the DCMI Abstract Model [Draft]

  “The Dublin Core Abstract Model requires that all terms (elements, element refinements, encoding schemes and controlled vocabulary terms) … that are compliant with the model must be assigned a URI reference that identifies the term.”

  Andy Powell, “Guidelines for assigning identifiers to metadata terms”
  <http://www.ukoln.ac.uk/metadata/dcmi/term-identifier-guidelines>
Vocabulary Term URI Construction

• Goals
  – *Persistence*. URI should have indefinite existence (in perpetuity?)
  – *Uniqueness*. By definition, URI should never identify more than one entity (vocabulary, vocabulary term or version of a vocabulary or vocabulary term)
  – *Resolvability*. If possible, URI should resolve using currently available Web browsers

• URI reference partitioning
  – XML namespace URI reference
    E.g., http://purl.org/gem/GEMS/
  – Vocabulary term name
    E.g., Physics
  – Resulting URI
    http://purl.org/gem/GEMS/Physics
Persistent URI Reference Strategies

Possible strategies:

- Use project-specific URL
  E.g., http://myproject.org/metadata/vocabs/color#Red
  • Questionable persistence
- Use PURL
  E.g., http://purl.org/gem/educationLevel/Grade 2
  • Reliable intermediary (resolution service) for persistence
- Use “info” URI
  E.g., info:ddc/22/eng//004.678
  • Persistent identification but info URIs cannot be “resolved” using current Web browsers

DCMI Working Draft: “Guidelines for assigning identifiers to metadata terms”
<http://www.ukoln.ac.uk/metadata/dcmi/term-identifier-guidelines>
Vocabulary Term Declaration

• **Assertion 2:** All controlled vocabulary terms must be “declared”
  – To “declare” a term is to create a machine-processable representation of the term by means of a schema language
  – XML and RDF/XML
  – Relationships among versions of a vocabulary and its terms should be declared

• Area of research and exploration
• See *Guidance information for naming, versioning, evolution, and maintenance of element declarations and application profiles* Draft CWA, July 2004
Examples: Declaration

• DCMI Type Vocabulary
• GEM Education Level Vocabulary
Vocabulary Management Information

<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE rdf:RDF (View Source for full doctype...)>
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
    <!-- CONTROLLED VOCABULARY MANAGEMENT INFORMATION -->
    <rdf:Description rdf:about="http://purl.org/dc/dcmitype/">
        <dc:title xml:lang="en-US">The DCMI Types namespace providing access to its content by means of an RDF Schema</dc:title>
        <dc:publisher xml:lang="en-US">The Dublin Core Metadata Initiative</dc:publisher>
        <dc:description xml:lang="en-US">The DCMI Type Vocabulary. Entries are declared using RDF Schema language to support RDF applications. The Schema will be updated according to dc-usage decisions.</dc:description>
        <dc:language xml:lang="en-US">English</dc:language>
        <dc:source rdf:resource="http://dublincore.org/usage/terms/" />
        <dcterms:requires rdf:resource="http://purl.org/dc/elements/1.1/" />
        <dcterms:isReferencedBy rdf:resource="http://purl.org/dc/terms/" />
        <dcterms:issued>2000-07-11</dcterms:issued>
        <dcterms:modified>2002-05-22</dcterms:modified>
    </rdf:Description>

    <!-- TERM DECLARATIONS -->
    ...
    ...
</rdf:RDF>
Vocabulary Term Declarations

<!-- TERM DECLARATIONS -->
<dcterms:DCMIType rdf:about="http://purl.org/dc/dcmitype/Collection">
  <rdfs:label xml:lang="en-US">Collection</rdfs:label>
  <rdfs:isDefinedBy rdf:resource="http://purl.org/dc/dcmitype/" />
  <rdfs:comment xml:lang="en-US">A collection is an aggregation of items. The term collection means
that the resource is described as a group; its parts may be separately described
and navigated.</rdfs:comment>
  <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class" />
  <dcterms:issued>2000-07-11</dcterms:issued>
  <dcterms:hasVersion rdf:resource="http://dublincore.org/usage/terms/history/#Collection-001"/>
</dcterms:DCMIType>

<dcterms:DCMIType rdf:about="http://purl.org/dc/dcmitype/Dataset">
  <rdfs:label xml:lang="en-US">Dataset</rdfs:label>
  <rdfs:isDefinedBy rdf:resource="http://purl.org/dc/dcmitype/" />
  <rdfs:comment xml:lang="en-US">A dataset is information encoded in a defined structure
(for example, lists, tables, and databases), intended to be useful for direct machine
processing.</rdfs:comment>
  <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class" />
  <dcterms:issued>2000-07-11</dcterms:issued>
  <dcterms:hasVersion rdf:resource="http://dublincore.org/usage/terms/history/#Dataset-001"/>
</dcterms:DCMIType>
...
</rdf:RDF>
GEM educationLevel Vocabulary (XML)

<?xml version="1.0" encoding="UTF-8" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
    xmlns="http://purl.org/gem/instance/level/" targetNamespace="http://purl.org/gem/instance/level/"
    elementFormDefault="qualified" attributeFormDefault="unqualified">
  <xs:annotation>
    <xs:documentation xml:lang="en">GEM controlled vocabulary for designating the education level of the
    audience for a resource. Created in 1996 by the GEM Consortium.</xs:documentation>
  </xs:annotation>
  <xs:simpleType name="Level">
    <xs:union>
      <xs:simpleType>
        <xs:restriction base="xs:Name">
          <xs:enumeration value="Preschool education"/>
          <xs:enumeration value="Kindergarten"/>
          <xs:enumeration value="Grade 1"/>
          <xs:enumeration value="Grade 2"/>
          ...
          <xs:enumeration value="Grade 11"/>
          <xs:enumeration value="Grade 12"/>
          <xs:enumeration value="Adult/continuing education"/>
          <xs:enumeration value="Higher education"/>
          <xs:enumeration value="Vocational education"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:union>
  </xs:simpleType>
</xs:schema>
Examples:
Instance Metadata

- Dewey Decimal Classification
- GEM Subject
DDC Instance in RDF

<?xml version='1.0' encoding='UTF-8'?>
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
  xmlns:rdfs='http://www.w3.org/2000/01/rdf-schema#'
  xmlns:dc='http://purl.org/dc/elements/1.1/'
  xmlns:dcterms='http://purl.org/dc/terms/'>
  <rdf:Description>
    <dc:subject>
      <dcterms:DDC>
        <rdf:value>930</rdf:value>
        <rdfs:label>History of the ancient world (to ca. 499 A.D.)</rdfs:label>
      </dcterms:DDC>
    </dc:subject>
  </rdf:Description>
</rdf:RDF>
GEM Instance in RDF (Value URI)

```xml
<?xml version='1.0' encoding='UTF-8'?>
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
    xmlns:dc='http://purl.org/dc/elements/1.1'/
<rdf:Description>
    <!-- GEM hierarchical subject: Arts--Photography -->
    <dc:subject rdf:resource='http://purl.org/gem/instance/
        subject/GEMS/arts_photography'/>
</rdf:Description>
</rdf:RDF>
```
Vocabulary Term Publication

• **Assertion 3:** All controlled vocabulary terms must be “published”
  – Web-addressable vocabulary registries
  – Registry complexity continuum:
    • *From* simple, human-readable HTML pages documenting vocabularies and vocabulary terms
    • *To* complex blends of human-readable pages and declared machine-addressable vocabulary schemes
  – Registry examples:
    • *Simple Registry*—
      Gateway to Educational Materials:
      <http://www.thegateway.org/about/documentation/gem-controlled-vocabularies/>
    • *Complex Registry*—
      DCMI Registry:
      <http://www.dublincore.org/dcregistry/>
Simple Registry: GEM
GEM: Student Grouping Vocabulary
GEM: Student Grouping Vocabulary (RDF)
Complex Registry: DCMI
**DCMI Type Vocabulary**

The Dublin Core Metadata Registry

*Promoting the discovery and reuse of metadata.*

---

**Browse the registry by classification type**

<table>
<thead>
<tr>
<th>Display</th>
<th>Terms Overview</th>
<th>Browse</th>
</tr>
</thead>
</table>

| http://purl.org/dc/terms/DCMIType |

**Label**: DCMI Type Vocabulary

**Definition**: A list of types used to categorize the nature or genre of the content of the resource.

**Description**: The DCMI Type Vocabulary provides a general, cross-domain list of approved terms that may be used as values for the Resource Type element to identify the genre of a resource.

**Is Defined By**: http://purl.org/dc/terms/

**PDF Type**: TypeScheme

**PDF Type**: Class

**See Also**: http://dublincore.org/documents/dcmi-type-vocabulary/

**Type**: encoding-schema

**Has Version**: DCMIType-002

**Issued**: 2000-07-11

**Modified**: 2002-06-15

**Controlled Vocabulary**: Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text
DCMI Type Vocabulary RDF

```
   <rdfs:comment xml:lang="en-US">A list of types used to categorize the nature or genre of the content of the
resource.</rdfs:comment>
   <dcterms:issued>2000-07-11</dcterms:issued>
   <rdfs:seeAlso rdf:resource="http://dublincore.org/usage/terms/history/#DCMIType-002"/>
   <dcterms:modified>2002-06-15</dcterms:modified>
   <rdfs:label xml:lang="en-US">DCMI Type Vocabulary</rdfs:label>
   <rdfs:type rdf:resource="http://purl.org/dc/terms/TypeScheme"/>
   <dcterms:description>The DCMI Type Vocabulary provides a general, cross-domain list of approved terms that may be used as values for the Resource Type element to identify the genre of a resource.</dcterms:description>
   <rdfs:transferred rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
</rdf:RDF>
```
DCMI Type Term: “Collection”
DCMI Type Term “Collection” (RDF)

```

  - <rdf:Description rdf:about="http://purl.org/dc/dcmitype/Collection">
    <rdfs:isDefinedBy rdf:resource="http://purl.org/dc/dcmitype/" />
    <rdfs:type rdf:resource="http://purl.org/dc/terms/DCMIType" />
    <dcterms:issued>2000-07-11</dcterms:issued>
    <rdfs:label xml:lang="en-US">Collection</rdfs:label>
    <rdfs:comment xml:lang="en-US">A collection is an aggregation of items. The term collection means that the resource is described as a group; its parts may be separately described and navigated.</rdfs:comment>
    <dc:terms:hasVersion rdf:resource="http://dublincore.org/usage/terms/history/#Collection-001" />
  </rdf:Description>
</rdf:RDF>
```
Questions?

Thank you for your attention

Stuart A. Sutton
sasutton@u.washington.edu
Joseph Tennis
jtennis@u.washington.edu