Vocabularies
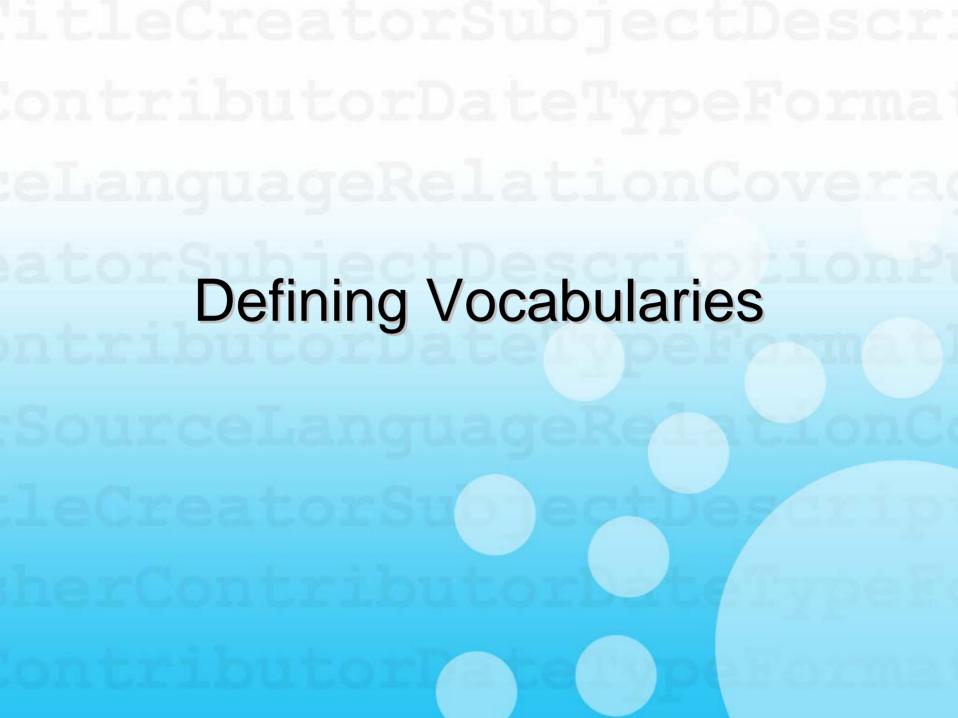
Vocabularies Tutorial
Manzanillo, Mexico
October 5, 2006

Joseph T. Tennis
The University of British Columbia

Outline

Vocabularies

1. Semantics:

   Defining, Developing, and Reusing

2. Posting to the web:

   Identifying, Declaring, Publishing, and Registering

3. Reuse on the web:

   Repurposing and Describing

Defining Vocabularies

Defining Vocabularies

Vocabularies

> A prescribed set of consistently used and carefully defined terms (DCMI Glossary)

ANSI/NISO Z39.19
- Lists
- Synonym Rings
- Taxonomies
- Thesauri

Defining Vocabularies

Examples:

Art and Architecture Thesaurus

<http://www.getty.edu/research/conducting_research/vocabularies/aat/>

NASA Thesaurus

<http://www.sti.nasa.gov/thesfrm1.htm>

Medical Subject Headings (MESH)

<http://www.nlm.nih.gov/mesh/>

DCMI Type Vocabulary

<http://dublincore.org/documents/dcmi-type-vocabulary/>

Defining Vocabularies

Vocabularies are made up of

Terms

Definitions (either by notes or by relationships or both)

A term from MESH:

Respiratory Therapy Department, Hospital

Definition:

Hospital department which is responsible for the administration of diagnostic pulmonary function tests and of procedures to restore optimum pulmonary ventilation.

Defining Vocabularies

Dublin Core glossary does not allow folksonomies to be vocabularies

For example: del.icio.us, flickr, connotea

This is because on the whole, they are neither consistently used, nor carefully defined.
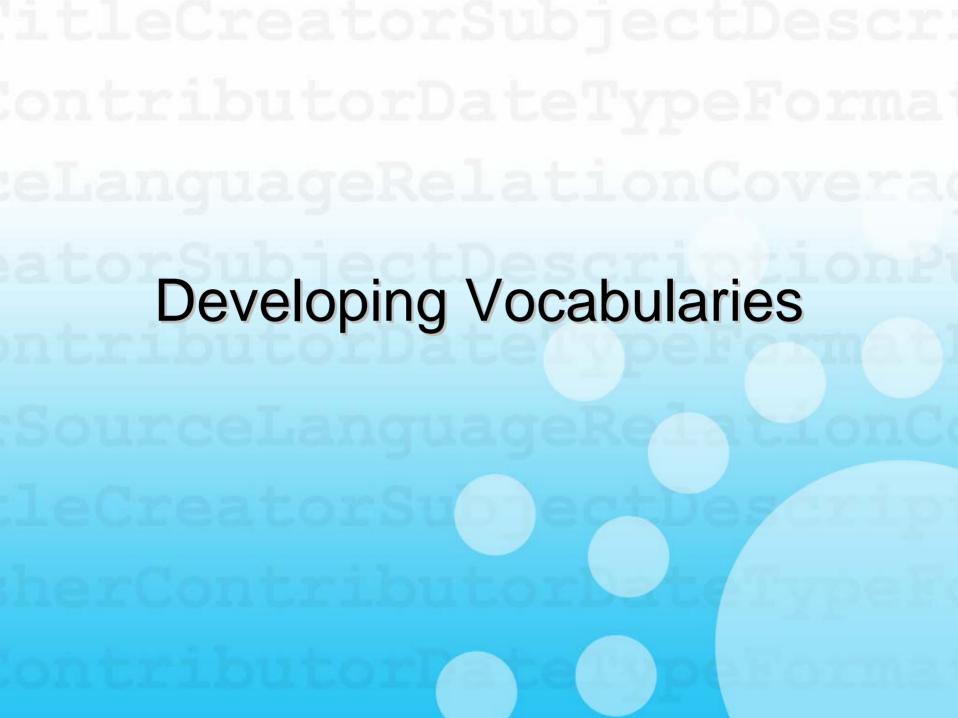
Defining Vocabularies

Network Environment

The Internet

Where humans and machines can link to other humans and machines

We want to design this linking so it is meaningful to all parties involved

This leads to recommending best practice for vocabulary specification and reuse in this environment

Developing Vocabularies

Developing Vocabularies

Observe the use of terms and control the meaning of concepts:

- Gather terms, concepts, and uses of those terms and concepts
- Document and make explicit the relationships between these terms,and concepts
- Make decisions about what to include and exclude based on use
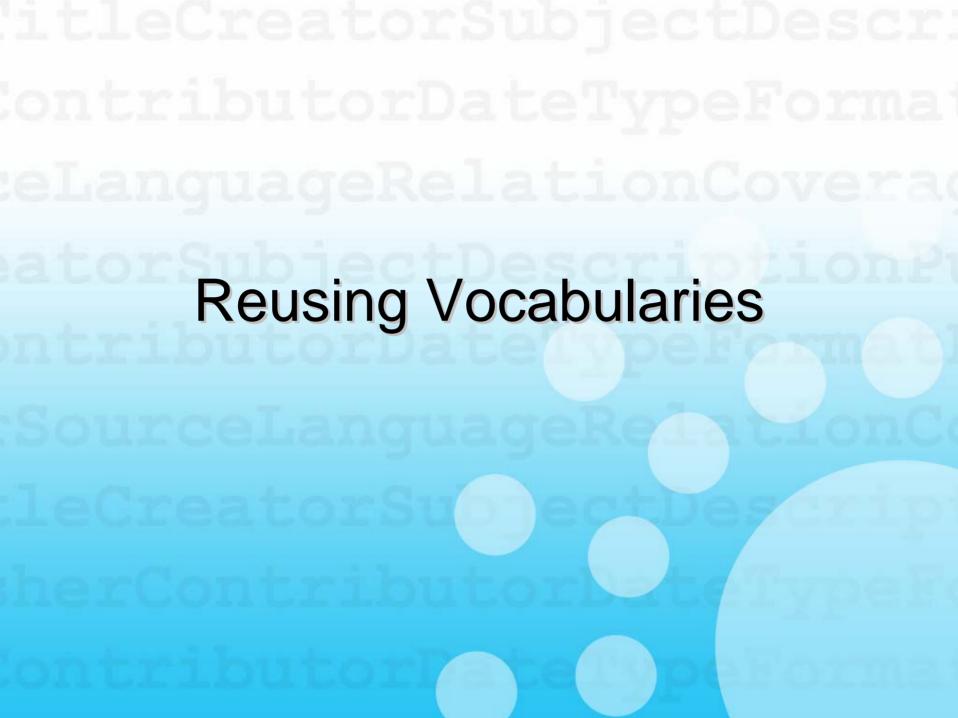- Value here is on decision to name and exclude some things

Developing Vocabularies

For more on developing vocabularies see:

Aitchison, Gilchrist, and Bawden's 2000 "Thesaurus Construction and Use: A Practical Manual"

and

Z39.19-2005

ANSI/NISO z39.19-2005 Guidelines for the Construction, Format, and Maintenance of Management Controlled Vocabularies
<http://www.niso.org/standards/resources/Z39-19-2005.pdf>

Reusing Vocabularies

Reusing Vocabularies

Vocabularies are the result a huge amount of effort, and if they are *owned* by an institution, then it is updated and maintained.

If vocabularies are available at addressable and machine processable parts of the networked environment we can facilitate reuse

Reusing Vocabularies

Vocabularies can be reused if:

You have permission to reuse them

Machines can reuse vocabularies if they are:
identified (given a URI reference),
declared (machine processable),
published (web accessible),
registered (contextualized and maintained).

Vocabularies in the Networked Environment, DC 2006, Manzanillo, Mexico (c) Joseph T. Tennis

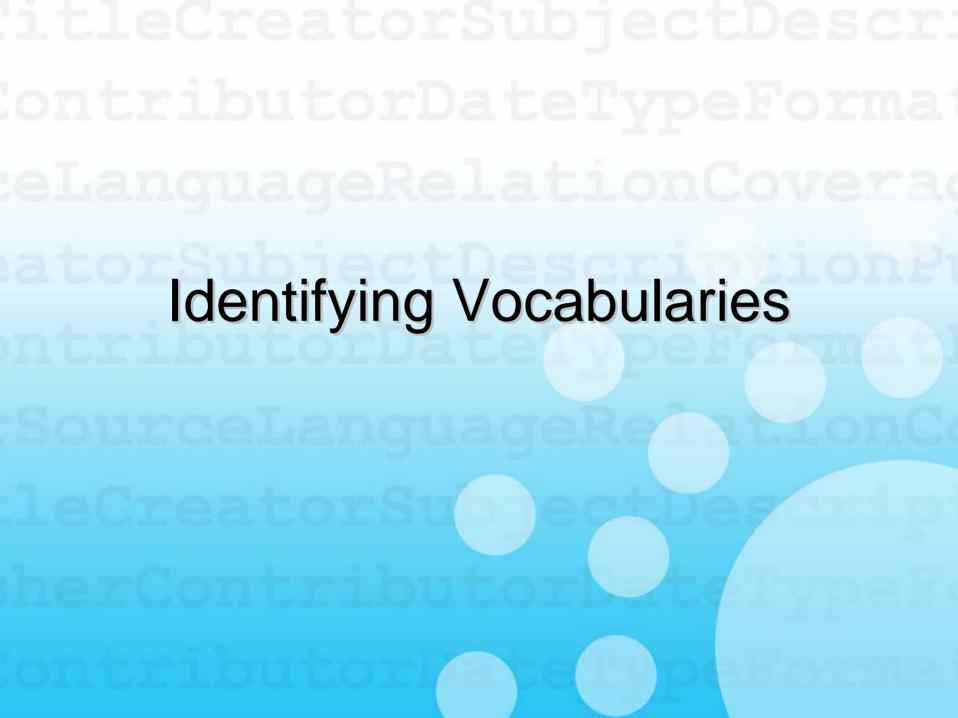# Recap 1

Recap 1

Defined

    vocabularies and networked environment

Developed

Reused

Reuse is the key to utilizing vocabularies to their full potential in the networked environment

Identifying Vocabularies

Identifying Vocabularies

"URIs identify resources and so are central to the Semantic Web enterprise. Using a global naming convention … provides the global network effects that drive the Web's benefits. URIs have global scope and are interpreted … across contexts. Associating a URI with a resource means that anyone can link to it, refer to it, or retrieve a representation of it."

Nigel Shadbolt, Wendy Hall, & Tim Berners-Lee, "The Semantic Web Revisited"
<http://eprints.ecs.soton.ac.uk/12614/01/Semantic_Web_Revisted.pdf>

Identifying Vocabularies

URIs are required by the DC Abstract Model

"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>

Identifying Vocabularies

Vocabularies contain terms

- Terms are resources
- They need to sit at a single space in the network
- They need a URI
- To that end, terms within a vocabulary need to be declared using a URI
- The URI should be persistent

Identifying Vocabularies

Strategies for identifying vocabularies and terms

Use project-specific URL

- E.g., http://myproject.org/metadata/vocabs/color#Red
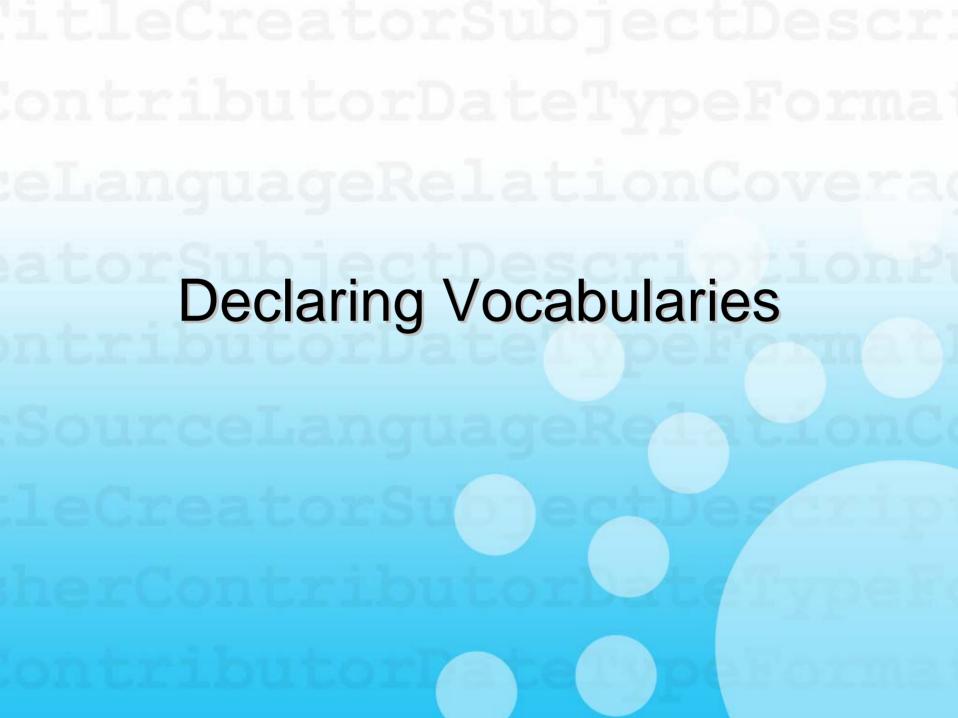
Questionable persistence

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>

Declaring Vocabularies

Declaring Vocabularies

Declaring a vocabulary in the networked environment means we create a machine processable representation of the vocabulary and its terms by means of a schema language

XML and RDF/XML

Declaring Vocabularies

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF ... >  …
<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">An aggregation of resources.</rdfs:comment>
<rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
<dcterms:issued>2000-07-11</dcterms:issued>
<dc:type
    rdf:resource="http://dublincore.org/usage/documents/principles/#vocabulary-
    term"/>
<dcterms:hasVersion
    rdf:resource="http://dublincore.org/usage/terms/history/#Collection-002"/>
</dcterms:DCMIType> …
</rdf:RDF>
```
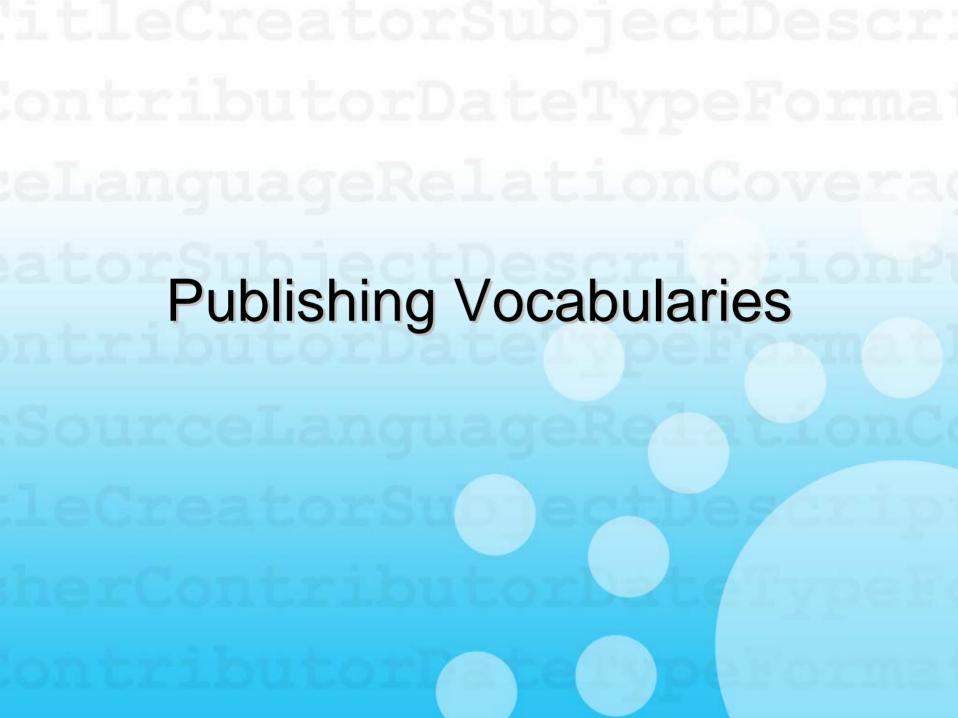
Declaring Vocabularies

```xml
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:dcmitype="http://purl.org/dc/dcmitype/"
    xmlns:dcterms="http://purl.org/dc/terms/" xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<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 Dublin Core Types namespace provides URIs for the entries of 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>
<dcterms:requires rdf:resource="http://dublincore.org/documents/dcmi-type-vocabulary/"/>
<dc:source rdf:resource="http://dublincore.org/documents/dcmi-type-vocabulary/"/>
<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>
<dcterms:TypeScheme rdf:about="http://purl.org/dc/terms/DCMIType">
<rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
<dc: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.</dc:description>
<dcterms:issued>2000-07-11</dcterms:issued>
</dcterms:TypeScheme>
</rdf:RDF>
```

Vocabularies in the Networked Environment, DC 2006, Manzanillo, Mexico (c) Joseph T. Tennis

Declaring Vocabularies

Versioning is an open issue at this point.

How do you make reference to outdated vocabularies and the most current vocabularies?

Publishing Vocabularies

Publishing Vocabularies

Vocabularies, once given a URI and a bound in a machine-readable schema, should be web-accessible.

This should be maintained by the owner(s) of that vocabulary

It should be give a URL

Should be resolvable and persistent

For example:www.myvocab.info
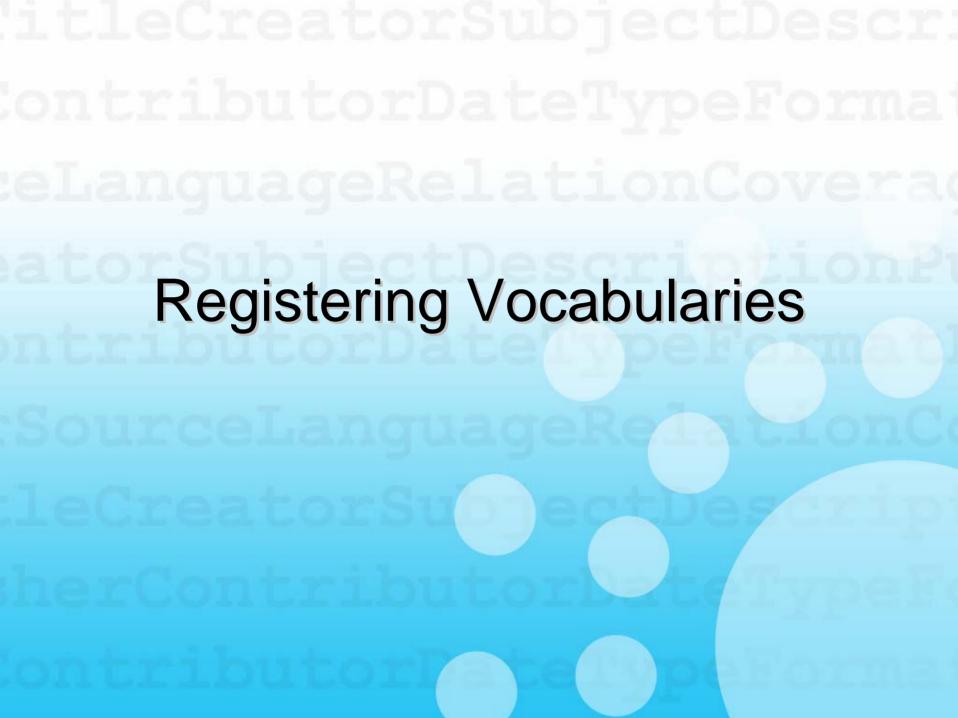
Publishing Vocabularies

For many it will be html pages to narrate the structure of the vocabulary.

However publishing is not just in .html, but also .xml or .rdf or .owl files offered through *content negotiation*

Where RDF/XML can be served to a machine

Fix for RDF/XML
<http://boakes.org/rdf-content-negotiation>
Content Negotiation
<http://httpd.apache.org/docs/1.3/content-negotiation.html>
Vocabularies in the Networked Environment, DC 2006, Manzanillo, Mexico (c) Joseph T. Tennis

Registering Vocabularies

Registering Vocabularies

Third party registries

Have a mandate to maintain published vocabularies from multiple parties

They require an explication of context (definitions, relationships, documentation, pointers to these, identification of owner(s) and editor(s)

Must commit to the requirements that come before registration: identification, declaration, and publication.
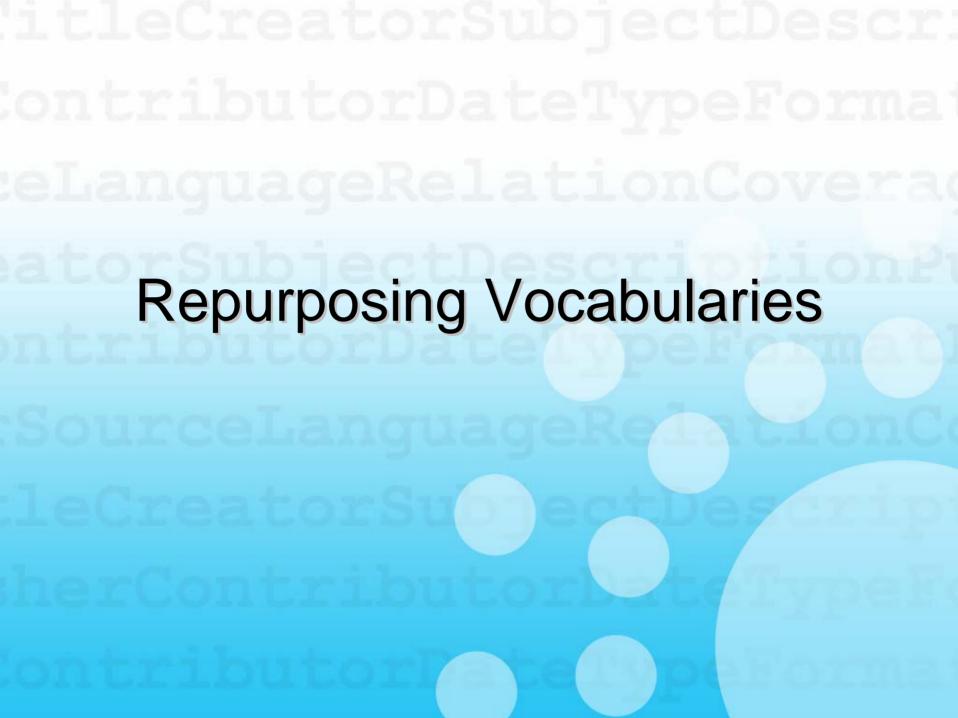
Registering Vocabularies

The value of registering vocabularies lies in the registry's ability to serve up versions of your vocabulary, contextualize your vocabulary, and maintain persistence (could help you identify, declare, and publish your vocabulary) - offering you complete networked vocabulary services

Recap 2

Recap 2

The networked environment is designed to help link humans and machines to humans and machines.

We can link humans and machines to vocabularies for human and machine use if we identify, declare, publish, and register vocabularies and their constituent terms.

Recap 2

Identify (through URI)

Declare (through machine processable representation)

Publish (through web accessible serving)

Registering (though submission to and contextualization in a third party server+services i.e., registry)
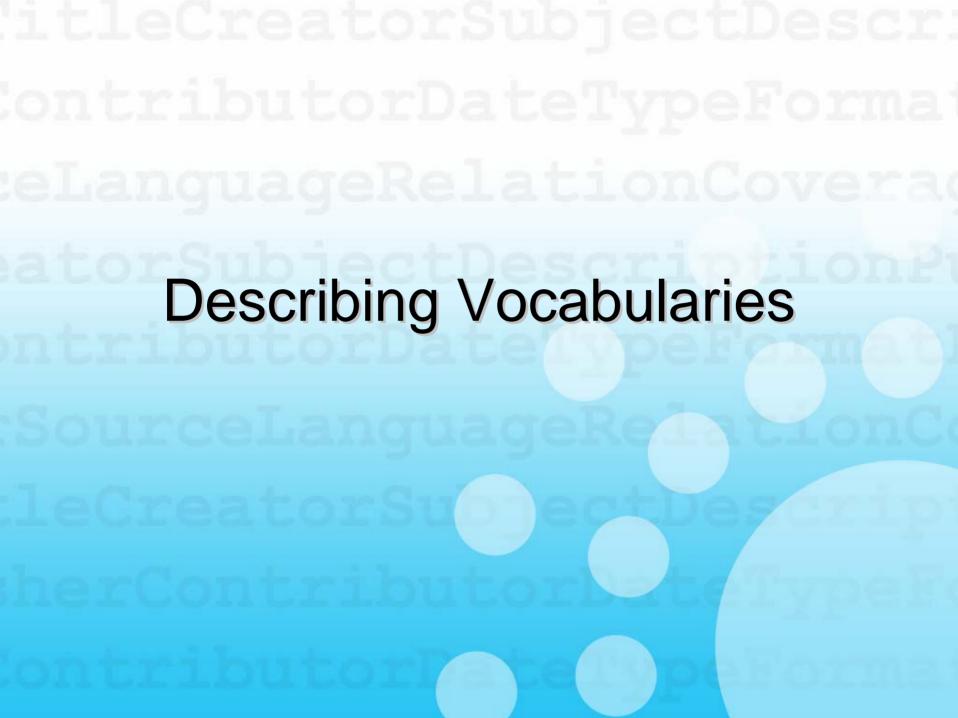
Repurposing Vocabularies

Repurposing Vocabularies

Once vocabularies have been registered, you can create repurposed vocabularies.

For example, you can repurpose a subset of DC Terms for your work.

You can also extend DC Terms to satisfy your needs.

Describing Vocabularies

Describing Vocabularies

Once we have vocabularies identified, declared, published and registered, we want to move them around the networked environment.

Describing Vocabularies

In order to do this we need to wrap metadata around the vocabulary describing it so we can make use of it in a different context, make relationships and definitions explicitly machine processable, map from one vocabulary to another, and identify differences between versions.

Describing Vocabularies

SKOS - Simple Knowledge Organisation Systems

w3c initiative

Lightweight specification for metadata about vocabularies

The purpose is to make meaningful assertions about vocabularies on the web

Describing Vocabularies

"SKOS is an area of work developing specifications and standards to support the use of knowledge organisation systems (KOS) such as thesauri, classification schemes, subject heading lists, taxonomies, other types of controlled vocabulary, and perhaps also terminologies and glossaries, within the framework of the Semantic Web."

<http://www.w3.org/2004/02/skos/>

Describing Vocabularies

Identifies

   Concepts (through URIs)

   Labels

   Relationships between concepts

   Change Notes & Scope Notes

Describing Vocabularies

SKOS is becoming less lightweight through community driven development.

They are wrestling with mapping, versioning, expressiveness, and other factors contribute to the expansion of SKOS.

Folks can contribute to this discussion.

Recap 3

Recap 3

Repurposing

The networked environment allows us to repurpose extant vocabularies in whole or in part.

Describing

In order to ship vocabularies around the networked environment with relationships and definitions intact, we must describe them in a standard way.

Documents

Defining, Developing and Reusing:

    ANSI/NISO z39.19-2005 Guidelines for the Construction, Format, and Maintenance of Management Controlled Vocabularies <http://www.niso.org/standards/resources/Z39-19-2005.pdf>

    Aitchison, Gilchrist, and Bawden 2000. "Thesaurus Consturction and Use: A Practical Manual" 4th Ed.

    Willpower Information Management Consultants <http://www.willpowerinfo.co.uk>

Identifying:

    RFC 3986 URI Generic Syntax 2005 <http://www.ietf.org/rfc/rfc3986.txt>

    Naming and Address: URIs, URLs… <http://www.w3.org/Addressing/>

Declaring and Publishing

    Expressing Simple Dublin Core in RDF/XML <http://www.dublincore.org/documents/2002/07/31/dcmes-xml/>

    Expressing Simple Dublin Core in XML <http://www.dublincore.org/documents/2003/04/02/dc-xml-guidelines/>

Registering

    ISO/IEC 11179 Metadata Registries <http://metadata-standards.org/11179/>

    Hillmann et al., 2006. "A Metadata Registry from Vocabularies Up: the NSDL Registry Project. <http://arxiv.org/ftp/cs/papers/0605/0605111.pdf>

Repurposing

    CWA 15249 Guidance information for naming, versioning, evolution, and maintenance of element declarations and application profiles <ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/MMI-DC/cwa15249-00-2005-Apr.pdf>

    DC Application Profiles Guidelines <http://dublincore.org/usage/documents/profile-guidelines/>

Describing

    SKOS <http://www.w3.org/2004/02/skos/>

Thank you
jtennis [at] interchange.ubc.ca

Acknowledgements:

Stuart A. Sutton, University of Washington

Diane Hillmann, Cornell University