Today’s Topics

- Metadata Overview
  - Standards
  - Some Definitions
- Web Metadata
  - Dublin Core Metadata Standard
  - Minnesota Web Metadata Standard
  - TagGen
Topics cont.

• Controlled Vocabularies / Thesauri
• A Look at Search Engines
  – North Star Search
  – Inktomi/Ultraseek Search Engine
• North Star Portal
  – Metadata application
  – Themes
Metadata Overview
What is Metadata?

• Structured data about data - the story
• Originally coined in the 1960s
• Library cataloging is one example
• Information on food packages is another
  – Vitamin content
  – Ingredients
Metadata Standards

• GIS: Geographical Information Service
• MARC: Machine Readable Cataloging, the library standard
• TEI: Text Encoding Initiative (SGML)
• Database fields
• Recordkeeping metadata
• Web metadata standards
What Does Metadata Do?

• Even good search engines need help
  – Better descriptions
  – Relevance ranking
  – Consistent terminology
What Else Does It Do?

- **Meta-meta**
  - Information about a resource adds value to resource discovery
  - Information about the metadata itself in the form of qualifiers
  - Fitness for use
Definitions
Some Definitions

- Metadata Standard
  - Standardized set of elements
  - Defined rules/options for use
- Metadata Registry
  - A registry of metadata standards
  - Defines elements within sets
More Definitions

• Metamodel
  – A formal model of metadata
  – Used in data warehousing
  – Also for db information exchange

• ISO 11179 Metadata Repository Standard
  – http://www.iso.ch/liste/JTC1SC32.html
  – Metamodell that describes the standard in a data structure that can be stored on a computer
"Doing research on the Web is like using a library assembled piecemeal by pack rats and vandalized nightly."

— Roger Ebert
Web Metadata
Web Metadata Options

• Garden-variety metatags
  – no control, no consistency in application
  – spammable, often ignored by search engines

• GILS
  – Government Information Locator Service
  – Federal standard - no longer supported
Dublin Core Metadata

- DCMI - Dublin Core Metadata Initiative
- DCMES - Dublin Core Metadata Element Set
- Started in 1995
- Stu Weibel - godfather of metadata
- W3C, librarians, and techies working together
D.C. Metadata Elements

- Title
- Creator/Author
- Contributors
- Subject/Keywords
- Description
- Publisher
- Dates: creation; last modified
- Identifier
- Resource type
- Format
- Relation
- Source
- Language
- Coverage
- Rights
<HTML>
<HEAD>
<TITLE>Foundations Project summary</TITLE>
<!-- Start Dublin Core - Do Not Modify Tags in This Block -->
<meta name="DC.Title" content="Foundations Project summary">
<meta name="DC.Subject" scheme="AACR2" content="Foundations Project (Minn.)">
<meta name="DC.Subject" scheme="LIV-MN" content="Environmental education">
<LINK REL=SCHEMA.DC HREF="http://www.bridges.state.mn.us/servlet/lexico">
<meta name="DC.Description" content="Summary information about the State of Minnesota's Foundations Project, including focus, purpose, design and intended results."">
<meta name="DC.Creator.PersonalName" scheme="AACR2" content="Quam, Eileen">
<meta name="DC.Publisher.CorporateName" scheme="AACR2" content="Minnesota. Dept. of Natural Resources">
<meta name="DC.Date.Creation" scheme="ISO 8601" content="1999-10-15">
</HEAD>
</HTML>

et cetera ...
TagGen Metatag Generator

- Simplifies addition of Dublin Core metadata
- Embeds tags directly into <head> portion of html doc
- Handles PDF files, or any others
- XML output capabilities
- Batch updating
What About XML?

• What is XML?
  – Separate content from presentation
  – XSL for the latter
• Content = title, author, date ….
  Metadata!
• DTD: Document Type Definition
  – Links standardized metadata elements like Dublin Core to document
Controlled vocabulary / Thesauri
Controlled Vocabularies

• Controlled vocabulary
  – Unambiguous language
  – No synonyms or false hits
  – Advantages of natural language covered by keyword searching
• Gateway for retrieving more detailed information
• Provides comprehensive, targeted access to important concepts in the literature
• Highly requested
Vocabulary Control Options

- Classification systems
- Authority files
- Controlled term lists
- Uncontrolled term lists
- Thesauri
Classification Systems

- Follow an outline of knowledge
- Used to put an object in a specific place - in the traditional classification system each item has a single location
- Used to shelve books in a library, e.g. Library of Congress numbers; relates to primary subject heading in online catalog
- Not usually natural language
Authority Files

- Sometimes called naming conventions
- Lists of terms in the preferred format for use
- Frequently have cross references
- Typically used for corporate names, personal names, or for geographic names
THESAURUS

- List of terms, cross-references, and scope notes
- Hierarchical nature
- Cross-references
  - Anticipate synonyms – “See” refs
  - Related concepts – “See also” refs
  - Broader, narrower, related terms
MN Thesaurus

- Legislative Indexing Vocabulary
  - Thesaurus used for MN web docs
  - Located at: [http://bridges.state.mn.us/servlet/lexico](http://bridges.state.mn.us/servlet/lexico)

- Simpler to use:
  - CCE topics
  - Located at: [http://search.state.mn.us](http://search.state.mn.us)

- Choose appropriate terms and place in Dublin Core subject element in TagGen
Mineral industries

Scope note: Use for technical and economic works on mining, metallurgy, and minerals of economic value. For works dealing with mining, use Mining engineering. For descriptive and statistical works, use Mines and mineral resources. For works dealing with metallurgy, use Metallurgy.

Used for
- Mines and mining
- Mining

Narrow term
- Ceramic industries

Broader term
- Industry

Related term
- Mines and mineral resources
Software That Embeds Vocabulary

- TagGen Metatag Generator
  - TagGen Office comes with several vocabularies, including:
    - Subject Tree developed by Washington State
    - Illinois State subject tree - other states adopting
    - General vocabulary
This screen is used to edit and create HSTF files. Use it to add, delete, sort, rename and move branches.

You can expand a branch with a [+], next to it by clicking the [+], sign, or collapse an expanded branch by clicking the [-] sign.
This screen is used to edit and create HSTF files. Use it to add, delete, sort, rename and move branches.

You can expand a branch with a [+] next to it by clicking the [+] sign, or collapse an expanded branch by clicking the [-] sign.
Search Engines
A Look at Search Engines

• Relevance ranking
• Descriptions
• Browseable categories
  – Offers controlled vocabulary
  – Examples
    • Yahoo
    • North Star Search: http://search.state.mn.us
North Star Search
http://search.state.mn.us

- Inktomi (formerly Ultraseek) is the State of Minnesota’s search engine
- 285 web sites indexed
- Nearly 400,000 documents
- Customizable search interface available to agencies
Marrying Controlled Vocabulary with Searching:

Inktomi’s Content Classification Engine
Content Classification Engine (CCE)

http://search.state.mn.us

- Yahoo-style directory
- Topic hierarchy
- Top-levels visible on North Star search
- Click on topic to see deeper levels
- Examples
Thesaurus and CCE

- North Star Search and Portal
  - LIV-MN is authority
  - CCE follows LIV-MN terminology, though less granular
  - WA-GILS subject tree spawned:
    - State of IL subject tree
    - These are similar to CCE topic layout
And Another Way to Embed Vocabulary

• Inktomi’s Content Classification Engine
  – State of MN search engine example: North Star Search
  – Follows state thesaurus (LIV-MN) terminology though less granular
  – Subtopics on discrete search interfaces: Bridges search
Searching and CCE

• Search by keyword
  – brings up relevant CCE topics
  – e.g. “nonpoint source pollution”
  – can use topics or look at results list
• Search by topic
  – start from North Star search
  – search by subtopic -- deeper levels of CCE
• Search within topic, by keyword
default within CCE
North Star Portal

http://www.northstar.state.mn.us
North Star Portal

• Themes
  – Business
  – Environment
  – Government
  – Health & Safety
  – Learning & Education
  – Living & Working
  – Travel & Leisure
North Star Portal

• Themes
  – Results linked from CCE topics

• Metadata template
  – DC-based
  – CCE topics pulled in for dc.subject selection
Web Metadata Guidelines
Web Metadata Guidelines
http://bridges.state.mn.us/bestprac/index.html

• Introduction & Overview: Minnesota Metadata Guidelines
• Recommended Metadata Standards & Software
• Training Manual
  – User Guide for Dublin Core Metadata
  – TagGen Download Information
  – TagGen Basic Instruction Guide
  – Visual Help Sheets
Web Metadata Guidelines (cont.)

- Bibliographies
- Appendices
  - Usability Studies
    - Dublin Core Metadata and Controlled Vocabulary on Bridges
    - Bridges User Interface
    - Needs Assessment
• Granularity
  – Add metadata to top page in folder (“index.html”)
  – Studies indicated effective metadata added to about top 50%
Web Metadata Guidelines (cont.)

• Appendices cont.
  – Reports
    • Thesaurus
    • Dublin Core Metadata Report
    • GILS Report
    • RDF/XML Report
• Comprises the State of MN Web Metadata Standard
• Training available
Metadata Overview

• Standards used
  – Metadata – Dublin Core
  – Controlled vocabulary/subject terms – LIV-MN or CCE
  – Name authority – state standard?
  – Geographic areas – library usage
  – Dates – ISO 8601
  – Language – ISO 639-1 (or ISO-639-2)
  – More … punctuation and capitalization
Software used
- Search engine – Inktomi
- Browsable categories – Inktomi’s Content Classification Engine
- Thesaurus – Lexico
- Metatag generation – TagGen or metadata templates
Summary

• Metadata is the key to resource discovery.
• Dublin Core aids search engines in discovery of web information.
• Use of controlled vocabulary important for specific yet encompassing search results, a powerful tool to access online electronic information.
• Consistent application of metadata requires research and application of standards.
Thank You!!

Eileen Quam
651.297.2341
eileen.quam@dnr.state.mn.us