Metadata Research and Application in The Management of Digital Collections

Laila Miletic-Vejzovic
University of Central Florida Libraries

Introduction

- Access to the online information
- Seeking information
- Web 1.0, 2.0, and 3.0
- Why metadata?
- The importance of metadata
- Use of metadata standards
- Metadata guidelines
- Implementation
- Additional resources

Access to the Online Information

- **Google and other search engines have revolutionized the way our patrons access information.**

- As per the commonly applied saying by the 21st century consumers in regard to any information, which actually applies perfectly to digitized material:

  " … if it's not [described] online, it does not exist."
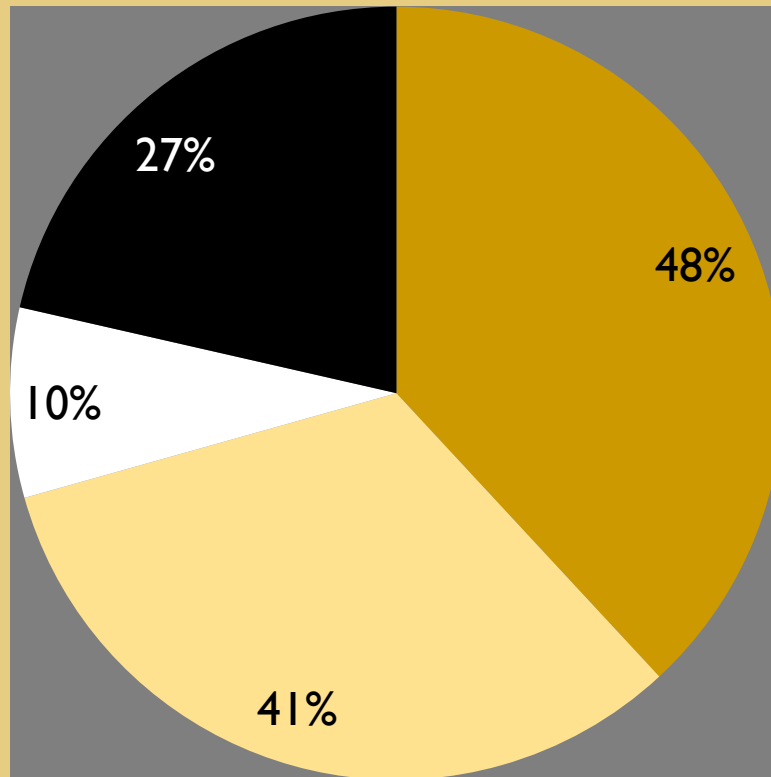
Seeking Information

Based on a report by Cathy De Rosa et al.:

Students started their searches using a search engine, i.e. Google:

- Only 48% ended up at a library Web site.
- Of which only 41% went on to use the library Web site.
- But only 10% indicated that the library Web site fulfilled their information needs.
- While 27% indicated that they had to use other resources.

De Rosa, Cathy et al., "College Students' Perceptions of
Libraries and Information Resources: A Report to the OCLC
Membership. A Companion Piece to the Perceptions of Libraries
and Information Resources," Dublin, Ohio: OCLC Online
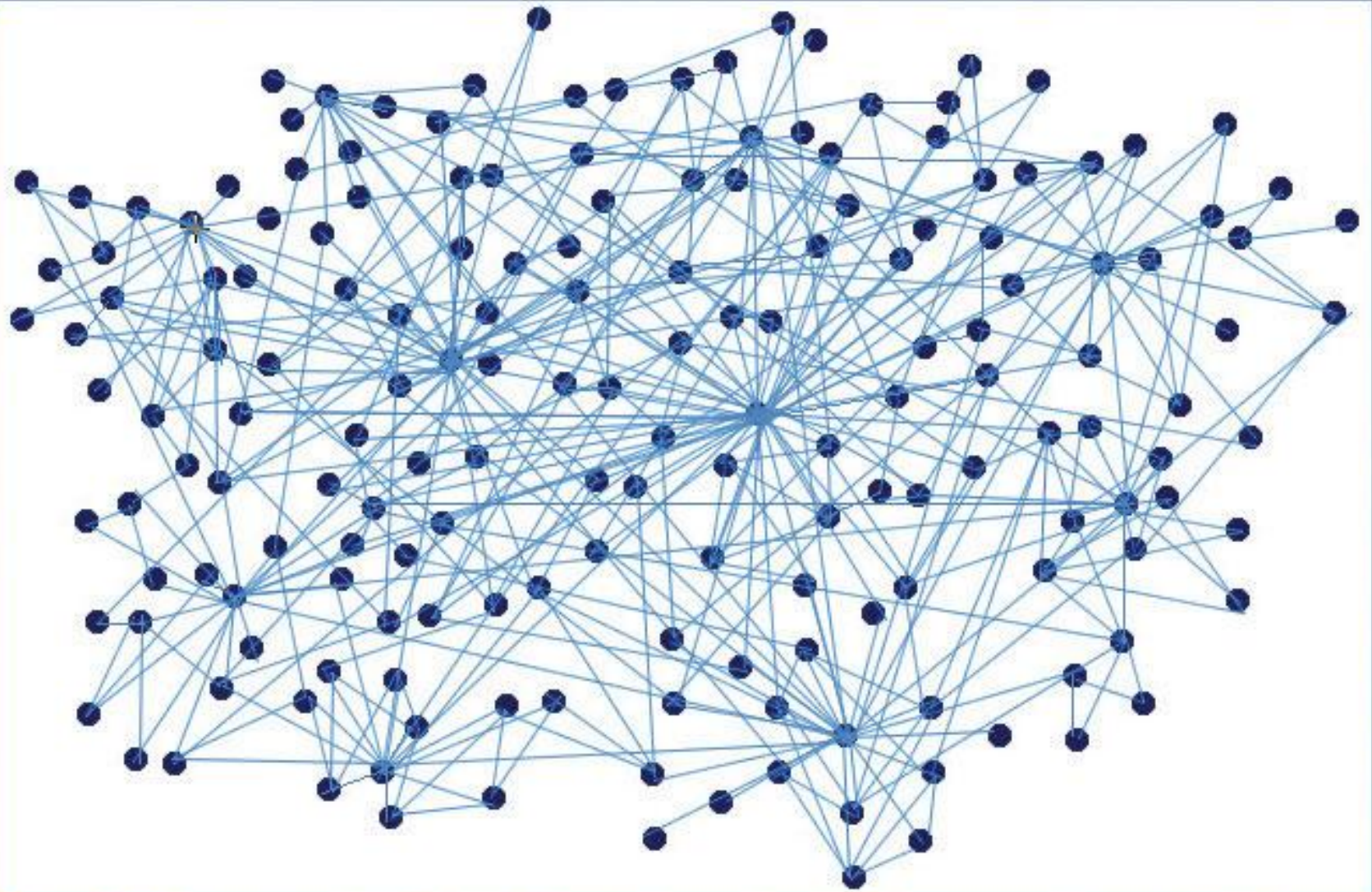Computer Library Center, Inc., 2006. [cited: 06-02]. Available at:
http://www.oclc.org/reports/pdfs/studentperceptions.pdf

Students started their searches using a search engine, i.e. Google

Pie chart:
- 48% Ended up at the library website
- 41% Went to use the library website
- 10% The library website fulfilled their needs
- 27% Had to use other resources

Internet

- Computers use network protocols to communicate in networks

- The Internet is a huge network of computers that use TCP/IP to communicate

- The Web is a system of interlinked documents accessed via the Internet

Figure 1: A figurative view of the World Wide Web. Users are allowed to link to any page they want but will naturally build communities of information with similar pages. The better sites receive votes of approval (in the form of links) from many pages.

The Evolution of Web

- Web 1.0 was the content Web –people could easily access the documents – which was the largest source of information ever

- Web 2.0 has generally been regarded as the social Web –people started to share photos and videos; interact in social networks; publish content in blogs; contribute in wikis; use tags and RSS – this is the read-write Web

W http://wikipedia.org/

Google

W Wikipedia

WIKIPEDIA

English
The Free Encyclopedia
2 942 000+ articles

日本語
フリー百科事典
600 000+ 記事

Deutsch
Die freie Enzyklopädie
928 000+ Artikel

Español
La enciclopedia libre
492 000+ artículos

Français
L'encyclopédie libre
825 000+ articles

Polski
Wolna encyklopedia
618 000+ haseł

Italiano
L'enciclopedia libera
585 000+ voci

Português
A enciclopédia livre
491 000+ artigos

Русский
Свободная энциклопедия
411 000+ статей

Nederlands
De vrije encyclopedie
546 000+ artikelen

search • suchen • rechercher • szukaj • 検索 • ricerca • zoeken • buscar • busca • поиск • sök •
搜索 • søk • haku • cerca • пошук • ara • hledání • keresés • căutare • suk • serĉu • søg • hľadať •
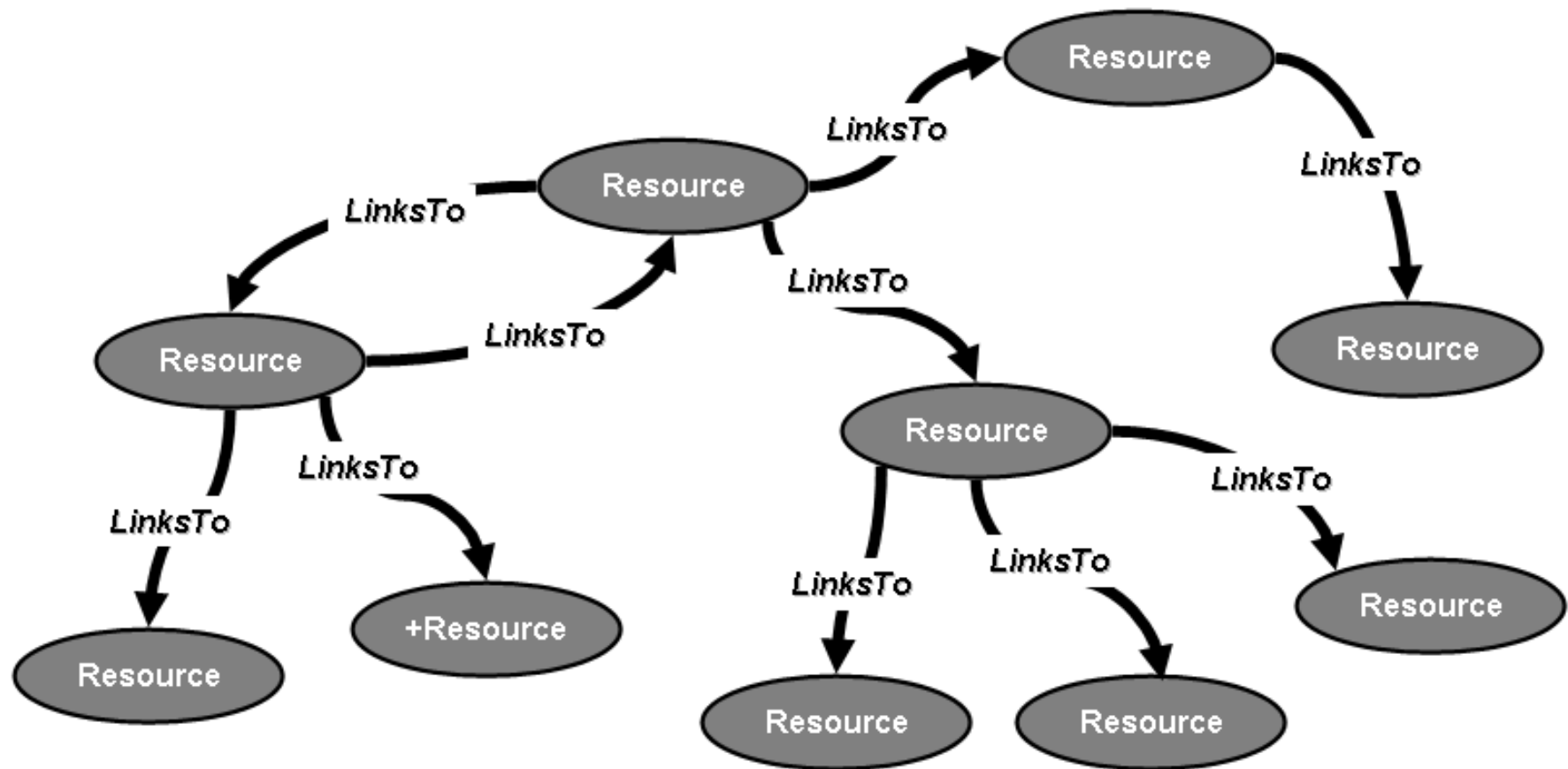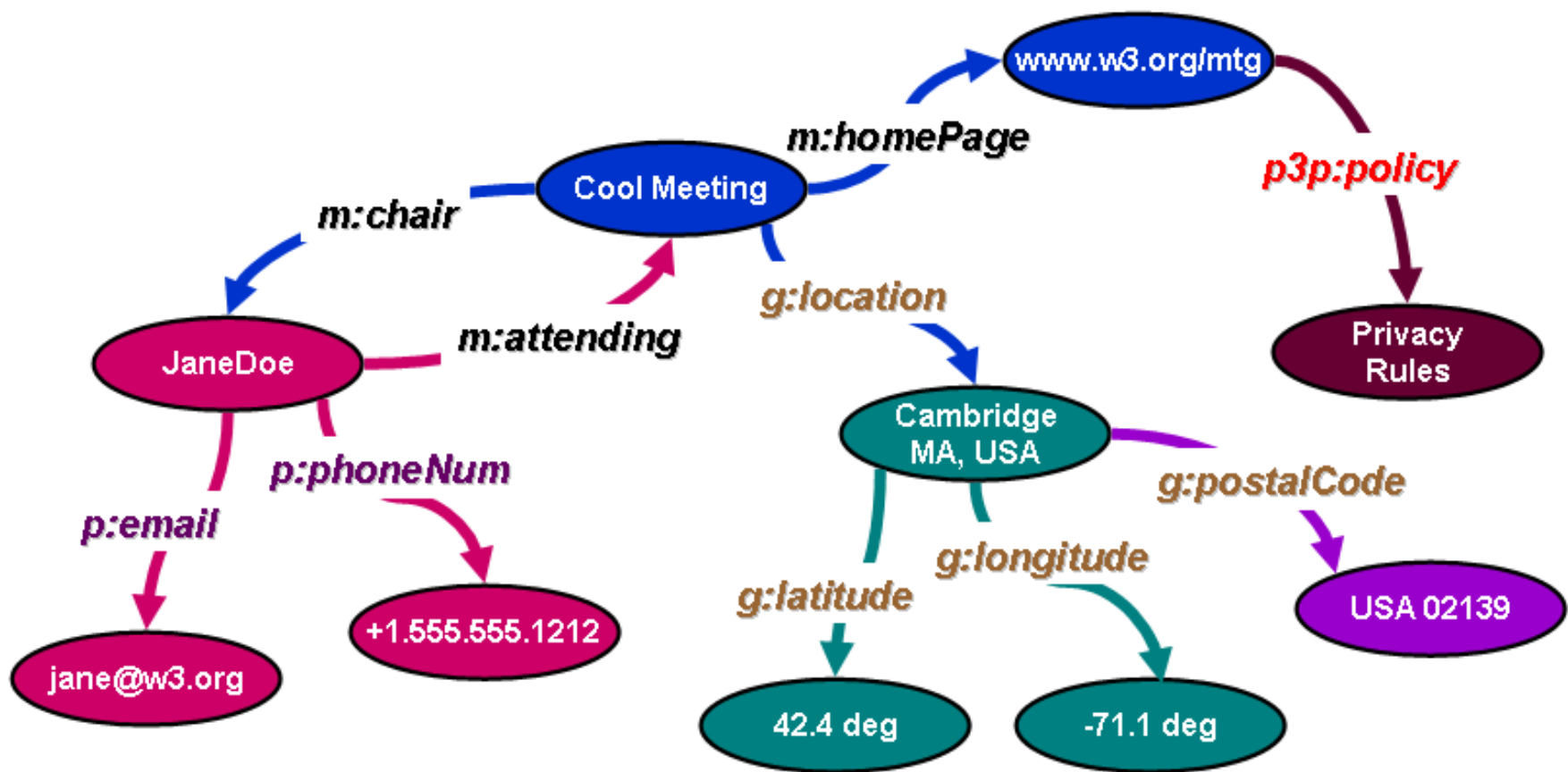cari • 검색 • بحث

English

→

Done

Web 3.0

- Web pages are written in HTML, which describes the structure of information

- If computers can understand the meaning behind information …

    - they can learn what we are interested in

    - they can help us better find what we want

- This is Web 3.0
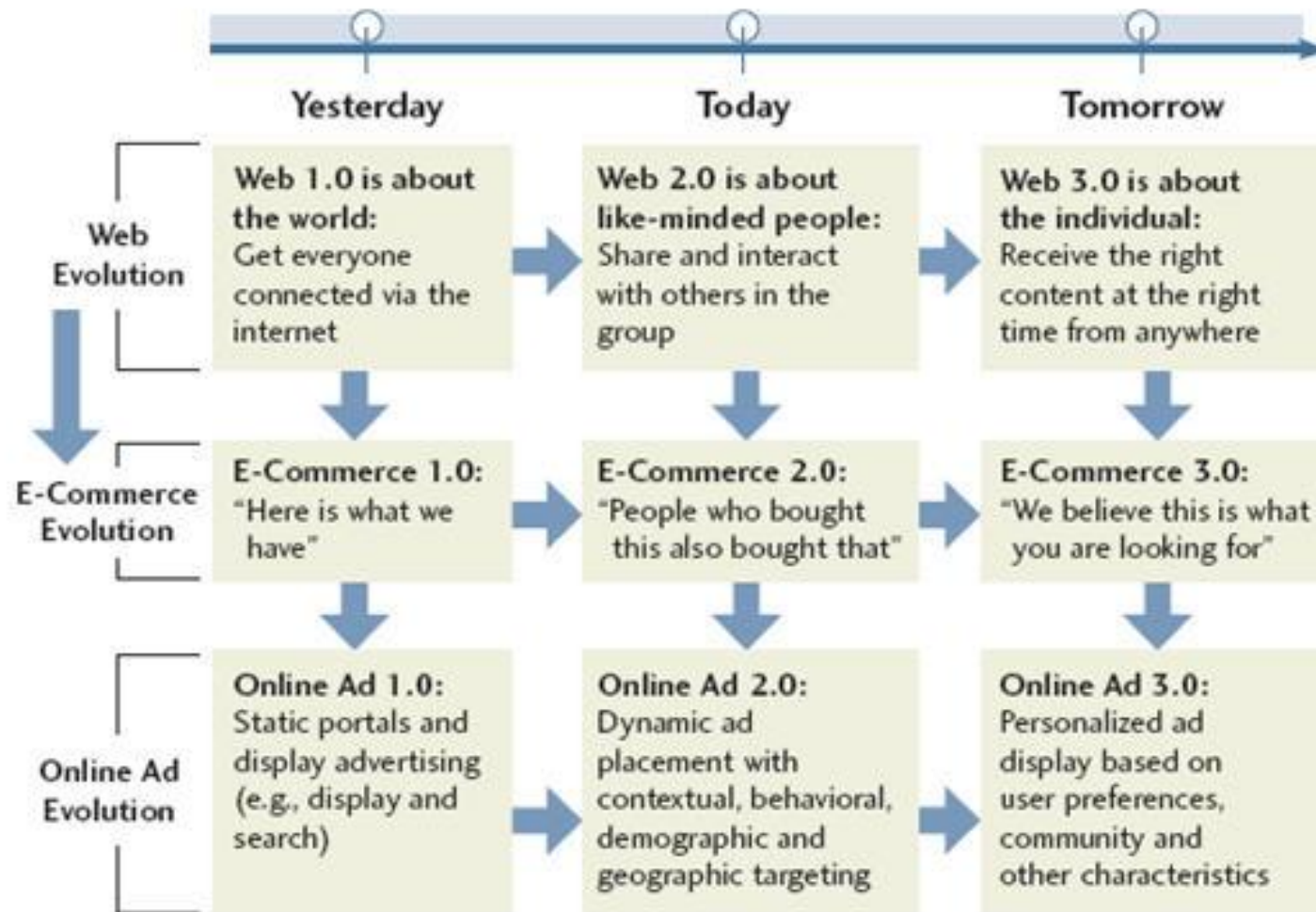
- Today's Web is about documents

Web 3.0 defined as:

- the location-aware and moment-relevant Internet

- the creation of high-quality content and services produced by gifted individuals using Web 2.0 technology as the enabling platform

- a third generation of Internet-based services, i.e. "the intelligent Web" (John Markoff of the NYT)

- a highly specialized information silos, moderated by a cult of personality, validated by the community, and put into context with the inclusion of metadata

Web Evolution and Its Impact on E-Commerce and Online Advertising

Source: Yankee Group, 2008

Descriptive Metadata?

- The primary purpose of descriptive metadata is to:

    - Assist researchers in the discovery of resources relevant to their research objectives.

    - Assist general and/or potential audience in discovery of resources relevant to their needs.

No Digitization Without Metadata

- Metadata is the backbone of digital curation.

- Metadata is essential!

- Without metadata a digital resource may be irretrievable, unidentifiable or unusable.

- Metadata is descriptive or contextual information which refers to, or is associated with, another object or resource.

Cont.

- Metadata consists of a structured set of elements which describe the information resource and assist in the identification, location and retrieval of it by users, while facilitating content and access management.

- Metadata standards formalize the element structure to ensure that the aims of a user community can be fulfilled.

Laila Miletic-Vejzovic
ACURIL XL

Basic Decisions about Metadata

Description:

- What kind of information do you need to describe each resource?

- What do your users need to know about what the resource is, where it came from, who created it, what its significance is?

- How much detail do you need to go into?

Cont.

Retrieval:

- How will users find resources in your collection?
- What will they be looking for?
- What aspects will they be interested in?
- At what level do you need to distinguish one resource from another, and at what level do you want to bring like resources together?

Formatting Data

Using standards for inputting your data is very important. Standards insure consistency, which:

- increases coherence and intelligibility of description
- enhances reliability of retrieval
- enables compatibility with other collections (cross-database searching)
- makes maintenance and possible migration of data easier

Cont.

Data should be formatted in a standard way. Actually, which format you choose may not be as important as always using the same format for data in the same field.

Examples:

- In a field called "Date" make sure that dates are always formatted in the same way.
- In a field called "Photographer" the same person's name should always appear in the same form.

- Similarly, the resources about the same topic should have the same term used to describe them.

- For example, a user looking for images of retail stores using the field "Subject" should be able to do a single search to find all the relevant images.

- If different terms are used, the user may not even realize that more than one search is necessary.

Controlled Vocabulary

- This is where a "controlled vocabulary" or "authority file" can be useful.

- A standard list of authorized terms can eliminate the ambiguity that arises from synonymous terms, homonyms, variant spellings and other pitfalls.

- There are controlled vocabularies that already exist for many subject areas and disciplines, or you could create your own standardized list of terms if it were reasonably short and you needed something very specialized for your collection.

Cont.

- Either way, with a controlled vocabulary you don't have to monitor your own consistency as you input metadata--the act of adhering to the list in itself will create the consistency you need.

- This is especially useful if more than one person will be inputting metadata in your collection.

Use of Metadata Standards

Benefits:

- Ensures rich, consistent metadata which will support the long-term discovery, use and integrity of digital resources.

- Ensures effective searching, improved digital curation and the possibility of sharing.

- Enables interoperability--metadata from a variety of sources can be integrated into other technical systems or machine read by compatible ones.

- The potential for resource discovery is much greater.

Interoperability

There have been many attempts at defining the concept of interoperability. A few examples are given below:

- "Interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality" [NISO, 2004].

- "Interoperability is the ability of two or more systems or components to exchange information and use the exchanged information without special effort on either system" [CC:DA, 2000].

- "Interoperability: The compatibility of two or more systems such that they can exchange information and data and can use the exchanged information and data without any special manipulation" [Taylor 2004, p. 369].

It is becoming generally accepted in the information community that interoperability is one of the most important principles in metadata implementation.

Types of Metadata Standards

- Effective implementation of metadata standards needs early consideration of the structure, content, functionality and links between digital objects and metadata instances required.

- Different types of metadata standards are used interdependently to achieve the following aims:

  - Metadata structure standards ensure consistent structure across individual entries; enable data searching to be implemented and data sharing across a discipline.

Cont.

- Hierarchical structure standards enable context as well as content to be described.

- Metadata content rules enable consistent data entry for effective searching.

- Content rules include: controlled vocabularies, authority files, thesauri, and classifications.

Searching CONTENTdm

- CONTENTdm has the capability to search multiple collections at once. In order to achieve this, CONTENTdm uses underlying mapping to simple Dublin Core (DC) elements to create a crosswalk between similar fields with different field names in different collections.

- The Dublin Core is an internationally agreed upon basic metadata scheme that defines 15 general descriptive elements, for example, Creator, Title, Date, Subject, Publisher, etc.

- You may map each field in your collection to a corresponding Dublin Core element. Or you could choose not to map certain fields to any DC element if the fields did not fit well into the DC schema, or if you didn't want to make these fields available for cross-database searching.

Example:

- The fields in the table below are from different databases and all somehow represent the name of a person (or organization) involved in the creation of a resource.

- Since all these fields have been mapped to the Dublin Core element "Creator", a cross-database search across multiple collections in the field "Creator" will retrieve the appropriate resources from whichever collection they are in, no matter what the collection-specific field name is.

Metadata Mapping Table

Collection	Collection-Specific Field Name	DC Mapping
Collection A	Architect	Creator
Collection A	Photographer	Creator
Collection B	Author	Creator
Collections C	Artist	Creator

Metadata

- Levels of metadata:

    - Determining the granularity, or detail, of metadata description is essential when developing a digital project.

- Types of metadata:

    - Descriptive: facilitates discovery and describes intellectual content.

    - Administrative: facilitates management of digital and analog resources.

    - Technical: describes the technical aspects of the digital object.

Cont.

- Structural: describes the relationship within a digital object.

- Preservation: supports long-term retention of the digital object and may overlap with technical, administrative, and structural metadata.

"Flattening Complex Reality"

- CONTENTdm's database structure right now is flat. There is no way structurally to distinguish between metadata for different physical manifestations of a resource, for example, between the original object, the photograph of the object, and the digitized scan of the photograph.

- The UCF Libraries has not attempted to follow a strict 1:1 correspondence between metadata and the particular manifestation of the resource.

- Whatever information seemed important for users of a particular collection was included in the metadata.

- For example, in a collection of photographs of buildings, both the photographer and the architect are important for searching, so both fields were included and both were mapped to the underlying Dublin Core element "Creator". The name of the person who did the scanning was not considered significant and was completely left out.

Setting up CONTENTdm Field Properties for Your Collection

You can set up your metadata fields in the CONTENTdm Server Administration module under "View/edit collection field properties." CONTENTdm allows you to:

- Have as many fields in the description as you want
- Create your own field names
- Decide whether each field will be searchable or will display
- Put the fields in any order you want
- Make fields available for cross-database searching

Field Properties Table

- To set field properties in CONTENTdm, use the Server Administration module, and select "View/edit collection field properties."

- Shown below are the default values for field properties as they appear in the CONTENTdm Server Administration module.

- *Remember*, the field properties as they originally appear in the Administration module are just a starting point--you can add, delete, and reorder the fields in any way, without affecting searching within the collection or across multiple collections. (It is the DC mapping that controls searching across multiple collections, not the order of the fields.)

Field name	DC mapping	Data type	Big field	Searchable	Hidden	Control Voc
Title	Title	Text	No	Yes	No	No
Subject	Subject	Text	No	Yes	No	No
Description	Description	Text	Yes	Yes	No	No
Creator	Creator	Text	No	Yes	No	No
Publisher	Publisher	Text	No	Yes	No	No
Contributors	Contributors	Text	No	Yes	No	No
Date	Date	Text	No	Yes	No	No
Type	Type	Text	No	Yes	No	No
Format	Format	Text	No	No	No	No
Identifier	Identifier	Text	No	No	No	No
Source	Source	Text	No	No	No	No
Language	Language	Text	No	Yes	No	No
Relation	Relation	Text	No	Yes	No	No
Coverage	Coverage	Text	No	No	No	No
Rights	Rights	Text	No	No	No	No

Conclusion

- Ensure metadata consistency and uniformity by developing:

  - Metadata guidelines for digitized material and collaborative projects among different institutions.
  - Best Practices for development of digital collections.

- Develop mapping guidelines for topical collections.
- Provide training opportunities.
- Train the trainers.

Thank you for your attention!

Questions?

For more information, please feel free to contact me at:

lmiletic@mail.ucf.edu

Appendices

A Framework of Guidance for Building Good Digital Collections, 3rd., Dec 2007

http://framework.niso.org/

A NISO Recommended Practice, prepared by the

NISO WG with support from IMLS

A FRAMEWORK OF GUIDANCE FOR BUILDING GOOD DIGITAL COLLECTIONS

COLLECTIONS

A digital collection consists of digital objects that are selected and organized to facilitate their discovery, access, and use. Objects, metadata, and the user interface together create the user experience of a collection. Principles that apply to good digital collections are:

❋ Collections Principle 1:

A good digital collection is created according to an explicit collection development policy.

❋ Collections Principle 2:

Collections should be described so that a user can discover characteristics of the collection, including scope, format, restrictions on access, ownership, and any information significant for determining the collection's authenticity, integrity, and interpretation.

Cont.

* Collections Principle 3:
  A good collection is curated, which is to say, its resources are actively managed during their entire life-cycle.

* Collections Principle 4:
  A good collection is broadly available and avoids unnecessary impediments to use. Collections should be accessible to persons with disabilities, and usable effectively in conjunction with adaptive technologies.

* Collections Principle 5:
  A good collection respects intellectual property rights.

* Collections Principle 6:
  A good collection has mechanisms to supply usage data and other data that allows standardized measures of usefulness to be recorded.

Cont.

- Collections Principle 7:
  A good collection is interoperable.

- Collections Principle 8:
  A good collection integrates into the users own workflow.

- Collections Principle 9:
  A good collection is sustainable over time.

Additional Resources

A number of excellent resources take a holistic view of digitization projects, covering topics ranging from selection, capture, and description to preservation and long-term access, and finally to metadata research. The following are highly recommended:

- UKOLN, Good Practice Guide for Developers of Cultural Heritage Web Services (2006) http://www.ukoln.ac.uk/interop-focus/gpg/

- Anne R. Kenney and Oya Y. Rieger, Moving Theory into Practice: Digital Imaging for Libraries and Archives (2000) http://www.library.cornell.edu/preservation/tutorial/ An online tutorial of imaging basics in English, French and Spanish

- Metadata Research Center <MRC> at School of Information and Library Science, University of North Carolina at Chapel Hill http://ils.unc.edu/mrc/

Cont.

- Northeast Document Conservation Center, Handbook for Digital Projects: A Management Tool for Preservation & Access (2000)

  http://www.nedcc.org/resources/digitalhandbook/dman.pdf

- Arts and Humanities Data Service (AHDS), Guides to Good Practice website
  http://www.ahds.ac.uk/creating/guides/index.html
  A series of guides to covering collection, description, and digitization for specific types of materials, such as GIS, performance resources, and virtual reality.

- Washington State Library, Digital Best Practices website
  http://digitalwa.statelib.wa.gov/newsite/digital/index.htm

- Susan Schreibman (editor), Best Practice Guidelines for Digital Collections at University of Maryland Libraries, 2nd ed. (2007)
  http://www.lib.umd.edu/dcr/publications/best_practice.pdf

Selected web site resources:

- North East Document Conservation Center (NEDCC): http://www.nedcc.org/home.php

- Society of American Archivists (SAA): http://www.archivists.org/

- Rare Books Manuscripts Section (RBMS): http://www.rbms.info/

- Association of Research Libraries (ARL): http://www.arl.org/

- National Archives and Records Administration (NARA): http://www.archives.gov/

- Library of Congress (LC):

  http://www.loc.gov/index.html

Selected Controlled Vocabularies & Thesauri available online

Library of Congress (LC) Subject Headings database:

http://id.loc.gov/authorities/

LC Authorities:

http://authorities.loc.gov/

LC Thesaurus of Graphic Materials I: Subject Terms (TGM-I):

http://www.loc.gov/rr/print/tgm1/

LC Thesaurus of Graphic Materials II: Genre & Physical Characteristic
Terms (TGM-II):

http://www.loc.gov/lexico/servlet/lexico?usr=pub-15:0&op=frames&db=TGM_II

Cont.

The Getty Thesaurus of Geographic Names (TGN):

http://www.getty.edu/research/conducting_research/vocabularies/tgn/

The Getty Art & Architecture Thesaurus:

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

The Getty Union List of Artist names:

http://www.getty.edu/research/conducting_research/vocabularies/ulan

Rare Books & Manuscripts Section (RBMS) Controlled Vocabularies:

http://www.rbms.info/committees/bibliographic_standards/controlled_vocabularies/index.shtml

Some Metadata Standards:

- Dublin Core (DC):
  http://dublincore.org/

- Metadata Encoding and Transmission Standards (METS):
  http://www.loc.gov/standards/mets/

- Metadata Object Description Schema (MODS):
  http://www.loc.gov/standards/mods/

- Describing Archives: A Content Standard (DACS):
  http://www.archivists.org/governance/standards/dacs.asp

- Encoded Archival Description (EAD):
  http://www.loc.gov/ead/

- EAD Help pages:
  http://www.archivists.org/saagroups/ead/