The Metadata Assignment and Search Tool Project

Anne R. Diekema
Center for Natural Language Processing
April 18, 2008, Olin Library 106G

SYRACUSE UNIVERSITY
School of Information Studies
INSTITUTE of Museum and Library SERVICES

Anne Diekema

- DEE-ku-ma
- Assistant Research Professor School of Information Studies – Syracuse University
- Interim Director of the Center for Natural Language Processing
- Research areas: library and information science, information retrieval, information organization, digital libraries, and natural language processing
- Collaborative applied research through funded projects
- MAST project funded by IMLS

INSTITUTE of Museum and Library SERVICES

SYRACUSE UNIVERSITY
School of Information Studies

Study Background

- Exponential growth of the WWW
- Importance of having digital presence beyond brick and mortar existence
- Development of digital collections
- Need efficient and effective means for making these collections searchable => metadata!
- Allows digital resources to be organized, validated and searched according to info not necessarily present in the full text of the digital resources such as dates, grade ranges, educational standards etc.

Metadata Assignment Woes

- Libraries limited by scarce funding and lack of expertise in cataloging digital collections
- One stumbling block is metadata assignment.
- Typically libraries copy catalog by downloading cataloging information from OCLC or Library of Congress
- Digital documents are not replications but most often original to the library holding it with very little metadata attached.
- The demands of creating metadata can be overwhelming to any library, but perhaps most noticeably to school and public librarians.

Cataloging Digital Collections

- Cataloging digital collections is tedious process
- High quality fully automatic cataloging not technically possible
- Most catalogers prefer hybrid approach
  - library administrators, catalogers, digital librarians, archivists
  - 70% or 148 participants preferred hybrid approach
  - only 1.4% insisted on fully automatic process (Greenberg, 2005)
- Computer-assisted cataloging tool that aids catalogers in assigning metadata to digital collections effectively and efficiently

INSTITUTE of Museum and Library SERVICES

SYRACUSE UNIVERSITY
School of Information Studies

MAST Project Goals

- Use of NLP to populate as many fields of the metadata record as possible
- Leave intellectual efforts of cataloging to the librarian who can now approve, improve, or add entries, rather than starting from scratch
- Provide necessary workflow support to aid the cataloger in efficient and accurate metadata generation.
- Disseminate metadata held within this tool using web services that allow customized search and discovery within institution's website

INSTITUTE of Museum and Library SERVICES

SYRACUSE UNIVERSITY
School of Information Studies

Team Work

- Metadata Assignment and Search Tool (MAST)
- Integrates three digital library tools and services
- Allows semi-automatic cataloging of digital resources
- Saves time, provides means for increasing availability of digital resources to underserved populations

- Center for Natural Language Processing (CNLP)
  - School of Information Studies, Syracuse University
- Digital Learning Sciences
  - University Corporation for Atmospheric Research (UCAR)

INSTITUTE of Museum and Library SERVICES

SYRACUSE UNIVERSITY
School of Information Studies

Digital Learning Sciences

- Digital Learning Sciences is a joint non-profit R&D center between UCAR (University Corporation for Atmospheric Research) and the University of Colorado Institute of Cognitive Science and Dept of Computer Science
- 22 staff, faculty, and graduate students – over 50 students involved since inception
- Interdisciplinary expertise in science education, educational technology, user-centered design and evaluation, cognitive science, computer science, computational linguistics, machine learning

INSTITUTE of Museum and Library SERVICES

SYRACUSE UNIVERSITY
School of Information Studies

DLS Activities

1. Develop tools, processes, and infrastructure for:
   - Using computational models of concepts to integrate, align, customize, and personalize
   - Accessing, organizing, managing, and enriching online resources
   - Operating digital libraries and repositories
2. Develop and curate collections and curriculum of online educational resources
3. Provide user and learner-centered design and evaluation services
4. Provide leadership, promote technology transfer, and engage in training and education in the above areas
5. Conduct research on how cognitive tools, computational algorithms, and interactive media can improve learning and engagement

Center for Natural Language Processing

- A multidisciplinary team of full-time information scientists, computer scientists, linguists, & educators
  - Strong academic credentials
  - Many with substantial commercial software experience
  - Immediate access to students/faculty with necessary expertise
- Research, development & licensing of NLP-based technology for government, industry, and foundations
  - NIH, NSF, NASA, NSA, DARPA, DTO, DOJ, DHS, NEH
  - ConEd, AT&T, MySentient, Unilever, ModSpec, SRC
  - Robert Wood Johnson Foundation, Mellon Foundation, OCLC
- Client-specific refinement of our NLP capabilities for very specific internal applications

INSTITUTE of Museum and Library SERVICES

SYRACUSE UNIVERSITY
School of Information Studies

Natural Language Processing

- Encompasses both theory and technology to enable a system to accomplish human-like understanding of text
- Recognizes explicit AND implicit meaning of text
  - Explicit : Entities, events, relations, time & topicality
  - Implicit: Subtler aspects of content, connotative meaning, certainty, affective, emotive, opinion & evaluative dimensions of meaning
- Disambiguates multiple senses of words / phrases based on global context (domain) and local context (meaning of surrounding words in sentences)

MAST to the rescue

- Tool addresses both quality and time issues with computer-assisted metadata assignment capabilities
- Uses Dublin Core and GEM elements
- Search service capability allows school and public libraries to provide a customized search mechanism that searches over the collections embedded within the institution's home website
- Support for educational standards assignment, an important focus of the United States educational system
- Standards metadata for teachers for easy adaptation of digital resources into their classrooms

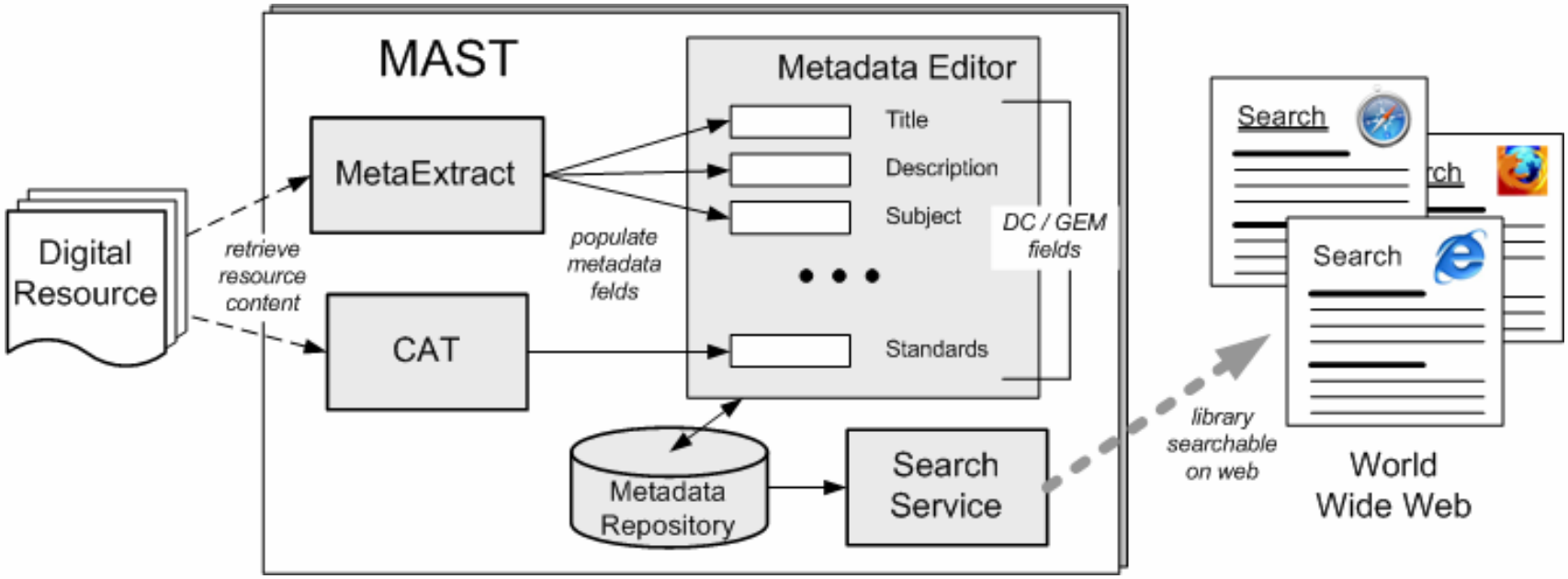
INSTITUTE of Museum and Library SERVICES

SYRACUSE UNIVERSITY
School of Information Studies

The Integrated Tool

- Combination of three different technologies
  1) Updated version of MetaExtract (Dublin Core + GEM)
  2) CAT, a computer-assisted educational standards metadata assignment tool
  3) Digital Collection System (DCS), a flexible tool that enables customized cataloging workflows and search capabilities
- Once integrated the tool will enable collection holders to quickly and richly describe their digital materials to make them fully available in a digital library or searchable via web services from their own website.

MAST System Diagram



INSTITUTE of Museum and Library SERVICES

SYRACUSE UNIVERSITY
School of Information Studies

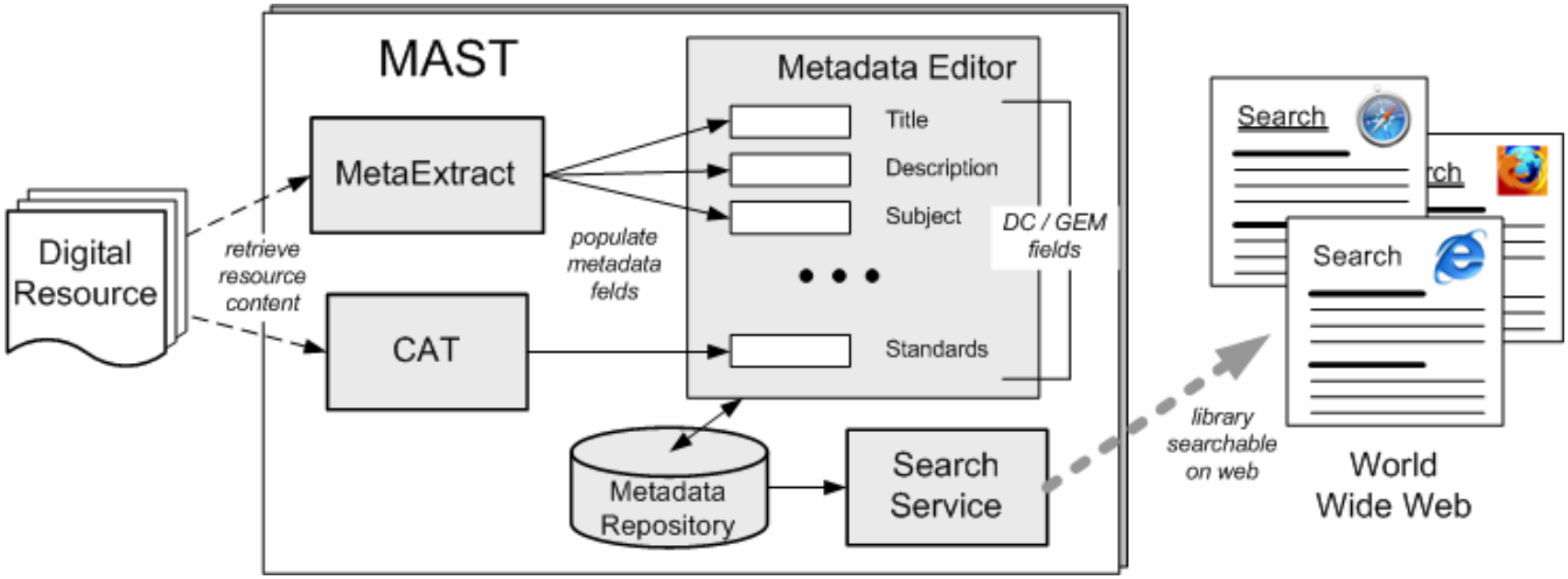
Digital Collection System

- Web-based cataloging tool that combines a metadata editor with search, discovery and OAI-PMH dissemination services
- Able to accommodate multiple metadata frameworks that describe items, collections, annotations and the like
- Frameworks must be expressed as an XML schema, which the tool uses to build the interface and enforce controlled vocabularies and required fields.
- A single instance can support multiple collections of different themes (topics) or metadata formats.

DCS Metadata Editor

- The metadata editor component of the DCS provides many supports for the cataloging process including:
- Ensuring well-formed records through schema-based validation of metadata values
- Ensuring required metadata fields have values
- Providing pick-lists for controlled vocabularies to eliminate typographical errors, and
- Providing best practices information that is easily accessible for each field.

MAST System Diagram



INSTITUTE of Museum and Library SERVICES

SYRACUSE UNIVERSITY
School of Information Studies

MetaExtract

- Component that extracts the bulk of the metadata values from the resource to populate the fields of MAST is MetaExtract.
- Information extraction system designed to extract metadata values
  - 15 elements of simple Dublin Core Metadata Element Set (DCMES)
    - Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights
  - 8 Gateway to Educational Material (GEM) elements
    - Audience, Cataloging, Duration, EssentialResources, Grade, Pedagogy, Quality, Standards

Extraction Modules

- MetaExtract compiles output from three distinct item-level extraction modules (and collection-level config file)
  - 1. TFIDF module
  - 2. Hbased module (extracts multi-sentence evidence)
  - 3. TextTagger rules (see slide that is coming up)
- Some elements are extracted from more than one module so that the system has a higher chance of populating the metadata elements
- Results of all of the modules are gathered and output as an XML file

MetaExtract and MetaTest

- NSF NSDL-funded research project
- Evaluated favorably in the MetaTest project (Liddy, 2005)
- No significant qualitative difference between automatic and manually assigned metadata
- Much better coverage of metadata elements
- Automatic metadata, manual, and full text perform comparably in retrieval and in quality for most elements
- Enables fielded searching, limits searches by particular aspects (e.g., grade, language), and allows easy browsing of results

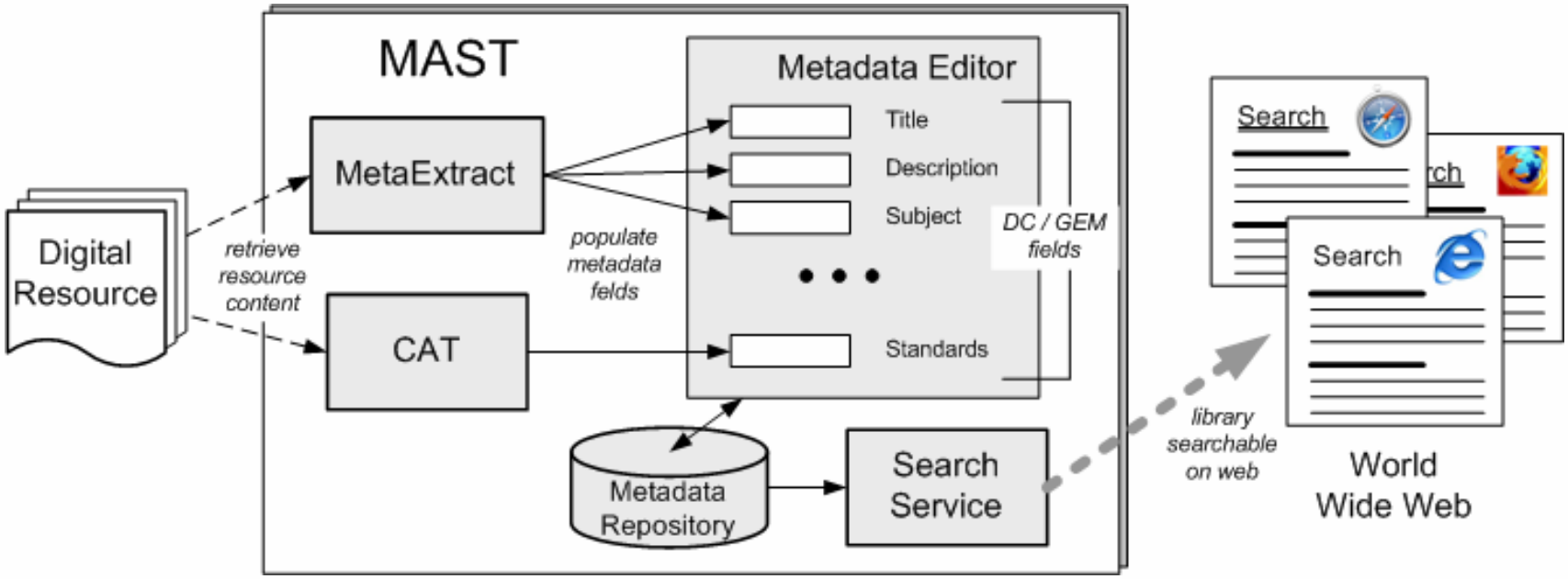
Metadata Record
Example

- Polygons (GEO0008.html)
- text = metadata
- audience-tool-for = Teacher
- pedagogy-teaching-method-process = play
- pedagogy-teaching-method-process = use
- **creator** = Marlena Mudryk
- date = May 10 , 1999
- grade = 4 , 5
- duration = one hour
- pedagogy-teaching-method = Demonstrations
- pedagogy-teaching-method = Discussions
- audience-beneficiary = Student
- publisher = AskERIC

INSTITUTE of Museum and Library SERVICES

SYRACUSE UNIVERSITY
School of Information Studies

MAST System Diagram



INSTITUTE of Museum and Library SERVICES

SYRACUSE UNIVERSITY
School of Information Studies

Content Assignment Tool (CAT)

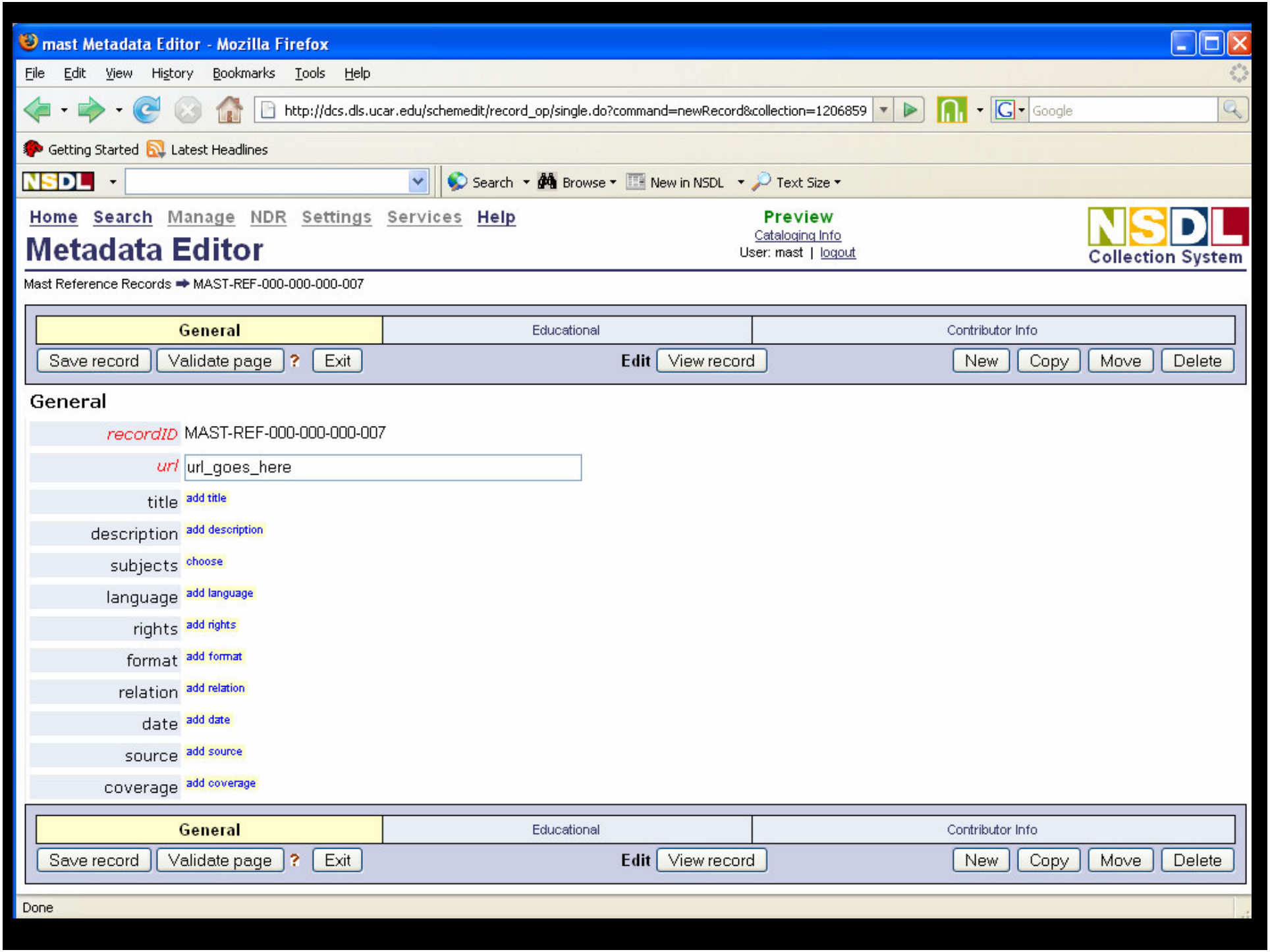
- NSDL Service project
  - to help speed up the standards assignment process, and
  - to provide a cross-walk between the different state and national standards
- Semi-automatic approach
- CAT suggests relevant standards to the cataloger
- Cataloger has final say in what gets assigned
- System learns from vetted assignments
- Audience: catalogers, curriculum and resource developers, teachers, school district administrators

INSTITUTE of Museum and Library SERVICES

SYRACUSE UNIVERSITY
School of Information Studies

The Proposed Metadata Schema

- New schema developed for this project
- Combination of OAI-Dublin core and GEM
- Three parts: general, educational, contributions
- 28 elements
- 2 from CAT
- 1 language is always set to English
- 1 cataloging done by cataloger not us
- 1 recordID
- 1URL
- This leaves 22 for MetaExtract (2 for CAT)


mast Metadata Editor - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://dcs.dls.ucar.edu/schemedit/record_op/single.do?command=newRecord&collection=1206859
Google
Getting Started Latest Headlines
NSDL
Search Browse New in NSDL Text Size
Home Search Manage NDR Settings Services Help
Preview
Cataloging Info
User: mast | logout
Metadata Editor
NSDL Collection System
Mast Reference Records ➡ MAST-REF-000-000-000-007
General Educational Contributor Info
Save record Validate page ? Exit
Edit View record
New Copy Move Delete
General
recordID MAST-REF-000-000-000-007
url url_goes_here
title add title
description add description
subjects choose
language add language
rights add rights
format add format
relation add relation
date add date
source add source
coverage add coverage
General Educational Contributor Info
Save record Validate page ? Exit
Edit View record
New Copy Move Delete
Done

System Evaluation

- CAT evaluations have been carried out previously
- MetaExtract will be updated and improved through this project
- To determine whether MetaExtract is performing correctly, and to try out various improvements, we propose to carry out a system-based evaluation which will
- Compare the metadata suggested by the system with assignments created manually by subject experts.
- System changes and improvements will be implemented based on this evaluation

User Evaluation

- A focus group at a major conference of library professionals
- Group of potential users to test the tools and provide input into functional specifications
- Combination of a design review and hands-on lab with guided tasks through the tool
- Gather data on the user experience
- Formal evaluation of the suggested metadata
- Discussion of the metadata provided by MetaExtract and CAT
- Participants will be asked to
  - postulate potential use cases in their home institutions or other settings
  - suggest modifications and additional functionality

Anne Diekema: diekemar@syr.edu