Abstract: At the NSDL Annual Meeting 2004, the SIG titled What's in the NSDL Metadata Repository? The Technical, Cataloging, and Evaluations Challenges at the NSDL All Projects Meeting will bring together technical, cataloging, and evaluation experts to identify methods to determine the breadth and depth of the metadata repository. During this session we will develop a plan to characterize the contents of the NSDL Metadata Repository in terms of subject, audience, and type. Please read the three attached documents before you come so that we will be able to make the most of our time and leave with a plan and some action items. Your help on this project will be greatly appreciated. The meeting will be at 10:30, Wednesday, November 17, 2004.

Report of the NSDL Evaluations Controlled Vocabulary Project
10/1/2004

Committee Members
Boots Cassel cassel@acm.org
Casey Jones caseyj@ucar.edu
Mimi Recker mimi.recker@usu.edu
Judy Ridgway Jridgway@enc.org (Working Group Chair)
Tammy Sumner sumner@Colorado.EDU

The tasks for the NSDL Evaluations Controlled Vocabulary Working Group were first identified in a teleconference meeting in February 2003. The two main tasks for this group were to create controlled vocabularies to assess the coverage of the NSDL collections and to facilitate the generation of evaluation reports. These reports would be used to visually represent the breadth and depth of the NSDL and would track how the collections are growing on an ongoing basis. In addition, the reports would help identify gaps in the collection so that those areas could be further developed.

The controlled vocabularies to assess the NSDL collection were limited to the metadata elements "audience", "discipline", and "type". Working group members took individual metadata elements and reviewed the metadata collected in the 2002 Evaluations Pilot Study. In addition, discipline specific vocabularies were used to address subjects not included in the pilot collections. As these evaluations vocabularies were being established, the group developed cross-walks to the metadata from the pilot study collections. It was determined that the element "type" would be best addressed though a combination of "format" and "media type". Each vocabulary was reviewed by two other working group members and revisions were made. The group posted the resulting draft vocabularies in the Educational Impact Standing Committee Workspace (http://eduimpact.comm.nsdlib.org/events/?pager=142) for public comment. The evaluations vocabulary cross-walks were tested against the vocabularies from nine additional collections.

Steve Bethard worked on using Lexical Analysis and SVM Classification to assess the holdings in the metadata repository. With the Lexical Analysis he matched words from subjects and audience to the *collection* level metadata and was able to identify 95.3% of the collections with at least one subject and 48.2% with at least one audience. This method does not use the description field to help identify any of the metadata. The remaining 5% of the collections were identified by hand and the lexical analysis was spot checked. Audience terms were problematic because they are not required terms.

Steve Bethard found the SVM Classification to be more flexible because it did take into account the data in the description fields. Another benefit that this method has over the lexical analysis is that SVM could interpret grade ranges. Janet Kahkonen Smith and Michael Flanagan cataloged and cross-checked 240 randomly selected records to generate training data for the SVM Classification. Once the enough training data was generated, Steve Bethard assessed both the collection level and item level metadata in the repository. Because many collections only supplied collection level records and no metadata for item level records, the exercise had poor results.

The results of this exercise were reviewed in February 2004. It was suggested that an approach that blended several methods be used. There was discussion about indexing the actual resources as well as using the provided metadata. Classifier and cluster approaches were discussed. To date, no action has been taken on these suggestions by the Collection Assessment Taskforce.

Once the assessment/reporting tool is in place, the working group recommends (1) that reports be generated at regular intervals, (2) the terms that do not fit into any of the cross-walks be monitored so that additional mappings can be developed as needed, and (3) a mechanism be developed such that the addition of new collections or changes in an existing collection's vocabularies can be supported as automatically as possible.

Metadata Elements in the NSDL Metadata Repository: Results of a Preliminary Quality Analysis: Subject, Type, Audience

Gregory M. Shreve, Marcia Lei Zeng, Bhagirathi Subrahmanyam

1. Introduction

In 2003 the NSF funded a targeted research proposal "Quality Analysis of Metadata Records in the NSDL Metadata Repository." The purpose of the targeted research was to take a snapshot of the metadata in the repository and to provide an analysis of metadata quality issues in the repository. Analysis parameters included the range, occurrence and distribution of Dublin Core elements in the repository as well as a profile of the correctness and consistency of assignment of content or values to metadata elements. Of special interest was determining levels of variation in metadata elements where standards could have easily been applied (e.g. ISO country-language codes for the DC *language* element).

The snapshot was taken in December 2003 after the last scheduled harvesting for that year. The methodology involved downloading the contents of the fifty-three collections in the repository at that time to our analysis site, using XSL stylesheets to process the downloaded OAI / NSDL xml files to extract the relevant metadata elements and place them into XML-based data files. XSL and server-side scripts were used to produce samples and statistical analyses. The count of valid records in the repository at that time was 186,237 records in fifty-three collections.

A look at table 1 gives some valuable insight into the repository. Without discussion of all of the implications of the statistics at this point, some preliminary conclusions are:

- Most of the collections do not use all the Dublin core elements available to them.
- Eight elements: *subject, creator, identifier, type, title, description, date* and *format* account for most of the elements in the repository.
- Some optional elements occur with greater frequency than those in the15 element core.
- Twenty-two elements, including three core elements, have frequencies below .01.

2. Subject, Type, Audience

With respect to the three elements under discussion in this SIG:

- *subject* element frequencies are relatively low over the collection. Collection average is approximately two subject elements per record. There are five collections that do not use a subject element.
- *type* occurs at the expected frequency, at or near 1 occurrence per repository record. Six collections of the fifty-three collections do not use the type element at all.
- *audience* is used relatively infrequently, in only six collections. Forty-seven collections did not use the element.

2.1 Subject

The quality analysis allows the following observations about *subject* (see Table 2):

1. The average number of *subject* elements per record in a repository collection covers a wide range. The highest average number of subjects is 38.6 and the lowest is 0.
2. Twenty-five collections, almost half of the collections in the repository, have an average number of subjects per record less than three; seventeen collections have an average of less than 1.5. This raises questions of the ability of the *subject* element to appropriately represent content in these collections.
3. There is a wide variation not only in average number of *subjec*t elements per record, but in the number of "discrete subjects." The relationship between the number of *subject* instances (Instance Count = IC) and discrete subjects (Discrete Count = DC) could be indicative of a number of situations.
4. IC significantly greater than DC could indicate lack of granularity in classification of content (See WGBH site with IC= 247 and DC = 9). If IC = DC it may indicate a very "flat" concept structure, with little or no attempt being made to represent taxonomies or hierarchical relationships. The use of subjects across records to indicate hierarchy or other classificatory relationship such as category membership would cause repetition of *subject* content and result in numbers IC > DC. An example of this is the PRI collection where IC=250 and DC = 255.
5. A look at the *subject* reports [1] indicates other possible problems. Below is an example from the PRI Collection: This example raises questions about what appropriate subject content should be (e.g., keyword, taxon, standardized term).

> 410 to 360 million years ago; Click on the specimens to learn more about them; Trilobites from the Devonian Period

6. Other issues that arise in the collections involve lack of rules for canonical form (articles or no articles, capitalization, plurals or no plurals), apparent lack of use of standard vocabularies, thesauri, taxonomies or other tools, and lack of base-level quality control (spelling errors, spacing errors).
7. Many collections opted not to use discrete *subject* elements, but "concatenated" subject content within a single element, complicating processing issues. Here is an example from the OpenVideo collection. This may be an attempt to create a taxonomy.

> Advertising: Screen Ads;Home economics: Laundry;Starch;

2.2 Type

Table 3 shows the complete listing of all values provided for the *type* element in the NSDL metadata repository. The table shows a wide variety of contents. Some general observations about include:

[1] available at http://appling.kent.edu/NSDLMetadataQuality/subject/SubjectAnalysis.Asp

1. Failure to use NSDL metadata primer recommended values (DCMI Type List.) Some of these recommended values do occur (note the frequency of text and image, among others), but so do many others.
2. Variant type categories may come from other *type* value space schemata, but are most likely local creations.
3. Superordinate categories (*visual*) have been created. Is this to be a superordinate of *image* and *stillimage* from the DCMI list?
4. Subordinate categories and a syntax for indicating subcategories emerges: *visual:map*.
5. Lack of quality control at input (*dataset* vs. *data set*)
6. Appearance of alternatives to the DCMI list (*Data:Modeled dataset* vs. *dataset*) occurs.
7. Mixed content (*Technical-Report MSC:: 65U05 Numerical methods in probability and statistics*) often occurs.

2.3 Audience

Table 4 shows the complete listing of all values provided for the *audience* element in the NSDL metadata repository. As indicated before the *audience* element is not widely used in the repository. The table shows a wide variety of contents for the value of the element: more than would be expected. Some general observations include:

1. Apparent lack of standardization
2. Does not appear to be heavy use of controlled vocabulary.
3. Large number of instances of “singleton categories” (*University teachers and students*).
4. There is concatenation within a single element instance instead of discrete audience instances (*K-12 students and teachers; undergraduates; general public*).
5. Lack of quality control at input: *High School (9-12) vs. HighSchool (9-12)*.

Table 1 Frequency of occurrence of metadata elements

Element	Type	Count	Frequency
subject	DC 15 Element	386861	2.08
creator	DC 15 Element	354499	1.90
identifier	DC 15 Element	240402	1.29
type	DC 15 Element	198438	1.07
title	DC 15 Element	186184	1.00
description	DC 15 Element	185003	0.99
date	DC 15 Element	161157	0.87
format	DC 15 Element	146921	0.79
language	DC 15 Element	98251	0.53
rights	DC 15 Element	83128	0.45
publisher	DC 15 Element	70753	0.38
contributor	DC 15 Element	62372	0.33
abstract	DCT Optional	55496	0.30
audience	DCT Optional	32187	0.17
references	DCT Optional	30796	0.17
educationLevel	DCT Optional	27215	0.15
created	DCT Optional	20351	0.11
isPartOf	DCT Optional	12693	0.07
source	DC 15 Element	12654	0.07
extent	DCT Optional	10749	0.06
conformsTo	DCT Optional	8712	0.05
issued	DCT Optional	8460	0.05
relation	DC 15 Element	6273	0.03
coverage	DC 15 Element	5884	0.03
isReferencedBy	DCT Optional	1125	0.01
temporal	DCT Optional	732	0.00
mediator	DCT Optional	314	0.00
isVersionOf	DCT Optional	183	0.00
modified	DCT Optional	141	0.00
medium	DCT Optional	140	0.00
hasPart	DCT Optional	117	0.00
hasVersion	DCT Optional	91	0.00
required	DCT Optional	48	0.00
alternative	DCT Optional	22	0.00
isRequiredBy	DCT Optional	22	0.00
tableofContents	DCT Optional	15	0.00
spatial	DCT Optional	10	0.00
available	DCT Optional	7	0.00
replaces	DCT Optional	2	0.00

Table 2 Collections sorted by # of subject elements per record

NAME	INSTANCES	AVERAGE	RECORDS	DISCRETE SUBJECTS
EARTHSCAPE	4285	38.60	111	548
INFOMINE	2489	30.73	81	1830
ESE REVIEWED	1171	23.42	50	309
EARTH SCIENCE	13929	18.67	746	1961
DISCOVER OUR EARTH	107	15.29	7	39
ICON	20354	8.60	2368	170
ENC	22649	8.43	2688	1421
COMET	45976	8.06	5703	3558
DWEL	947	8.03	118	225
AVC	106	7.57	14	33
GREEN	911	6.70	136	589
CUTTING EDGE	773	6.61	117	158
EARTH EXPLORATION	96	6.40	15	24
DLESE	24211	6.29	3847	1113
GENDERSCIENCE	2912	5.73	508	163
STARTING POINT	262	5.70	46	38
LEARNING MATRIX	5310	5.61	946	677
EUCLID	30335	4.75	6380	13941
NSDL NSDL	889	4.17	213	390
JCE DIGITAL LIBRARY	156	3.71	42	38
ALSOS	2165	3.50	618	139
MATHWORLD	35128	3.02	11645	545
DLNET	3237	3.00	1079	328
FEOL	11048	2.95	3739	4392
DSPACE	7134	2.13	3357	5028
CAV200	221	1.99	111	21
MATH FORUM	36606	1.95	18757	219
ALEXANDRIA	19766	1.75	11309	17
MATH-DL	45	1.67	27	26
GROW	708	1.64	431	103
LON-CAPA	21777	1.54	14112	8384
PREL	108	1.19	91	47
PLANETMATH	1984	1.16	1717	560
ARXIV	62080	1.12	55500	321
WGBH	247	1.03	240	9
CALTECH CSTR	414	1.00	414	2
CALTECH ETD	377	1.00	377	22
NSDL CEMAGAZINE	33	1.00	33	33
CALTECH EERL	293	1.00	294	5
PRI	250	0.98	255	250
NSDL SCIENCE ZONE	46	0.96	48	46
NSDL VIRTUAL TRAINING	19	0.95	20	19
OPENVIDEO	1748	0.88	1983	1072
INFORMEDIAVIDEO	466	0.43	1075	25

NSDL AWESOME	674	0.36	1865	526
BIOMED CENTRAL	2370	0.25	9597	2352
ECONPORT	46	0.19	237	17
ASDL	3	0.00	708	3

Table 3 Type elements in the metadata repository

Element Name	Count
Text	97729
Image	18662
InteractiveResource	11336
aerial photographs	8457
remote-sensing images	8454
Collection	4348
Text:Reference	3072
Maps	2852
Meeting abstracts	2493
Video	2014
Technical Report	1895
Research article	1763
encyclopedia entry	1717
Software	1691
Visual:Photograph	1684
Learning materials:Lesson plan	1398
Visual:Map	1253
Visual:Scientific illustration	1178
Learning materials:Classroom activity	1127
Research news	1045
Visual:Scientific visualization	984
Dataset	867
Visual:Remotely sensed imagery	777
Paper Report	764
Data:In situ dataset	741
Portal:Educational portal	722
Main	708
Research	672
Text:Report	654
Learning materials:Tutorial	627
Learning materials:Instructor guide	584
Research study	528
Review	500
Learning materials:Module or unit	497
Learning materials:Computer activity	495
Learning materials:Assessment	493
Book	486
Service	483
Text:Glossary	466
Service:Clearinghouse	413
Portal:Government portal	401
Visual:Video	400
Commentary	394

Learning materials:Lab activity	364
Text:Index or bibliography	327
Technical-Report	321
Data:Remotely sensed dataset	320
Image set	314
Article	308
Tool:Calculation or conversion tool	297
Text:Abstract or summary	287
Video	284
Tool:Software	259
Learning materials:Presentation or demonstration	217
image	217
Paper report	214
Visual:Artistic illustration	209
Learning materials:Course	204
Learning materials:Project	195
Portal:Nonprofit portal	193
Learning materials:Virtual field trip	191
Meeting report	179
Text:Periodical	175
Text:Journal article	171
Learning materials:Field activity	167
Learning materials:Case study	159
Text:Book	157
Data set	157
Meeting abstract	142
Sound	139
Learning materials:Field trip guide	133
Data:Modeled dataset	132
Service:Ask an expert	131
Minireview	129
Poster presentation	122
Methodology article	122
Reference	121
Working Paper	118
Service:Forum or discussion	118
Educator's guide	115
Case report	114
Conference Paper	111
Learning materials:Problem set	108
Service:Search engine	99
Web report	99
Secondary source	95
Conference or Journal Paper	88
Learning materials:Curriculum	86

Service:Listserv	86
Web Report	86
Environment	82
Document	75
Text:Policy or procedure	74
Learning materials:Syllabus	74
M.S./Diploma thesis	71
Website	68
Tool:Code	67
Offline:Physical object	56
Multimedia	51
Event	51
Editorial	50
Ph.D. thesis	49
Oral presentation	49
Preprint	47
Comment	47
Portal:Commercial portal	43
Debate	42
Audio:Sound	39
Tool	39
Thesis	39
Protein family review	38
Primary research	37
Opinion	35
Speaker presentations	33
POSTERS	33
Activity	33
Supplement Review	33
Text:Proceedings	31
Other	31
Community	31
Deposited research article	30
Habilitation	29
Study protocol	27
Service:Message board	27
Hypothesis	26
Supplement	25
EC-deliverable	24
Audio:Lecture	23
Letter	21
Book report	19
Audio:Audio webcast	18
Visual:Visual webcast	17
Film	17

Correction	16
Project-Report	15
Audio:Sound	15
Correspondence	14
Table	14
Methodology	13
Viewpoint	12
Audio:Music	12
Non-peer-reviewed research	12
Lecture Notes	11
Audio:Radio broadcast	9
Curriculum support	9
Article selection	9
Original investigation	9
Technical Report ORG-ID:: dilico.uhamburg_cs.TGI	9
Tutorial	9
Message	8
Method	8
Announcement	7
Audio:Oral history	7
ORAL PRESENTATIONS - SESSION 1	7
ORAL PRESENTATIONS - SESSION 2	7
ORAL PRESENTATIONS - SESSION 3	7
Original research	6
ORAL PRESENTATIONS - SESSION 6	6
Proceeding	6
Technical Report ORG-ID:: dilico.uhamburg_cs.TECH	6
Abstract	5
data	5
Short paper	5
Report ORG-ID:: dilico.uhamburg_cs.TGI	5
ORAL PRESENTATIONS - SESSION 5	5
ORAL PRESENTATIONS - SESSION 4	5
Focus	4
Case control study	4
Technical Report ORG-ID:: dilico.uhamburg_cs.KOGS	4
CDROM	4
Technical-Report LOCLANGUAGE:: Italian	4
Erratum	4
Technical Report ORG-ID:: dilico.uhamburg_cs.DBIS	4
Technical Report ORG-ID:: dilico.uhamburg_cs.SWT	4
Perspectives	3
Proceedings	3
Survey/Cross sectional study	3
Monography/Book chapter	3

remote-sensingimages	3
Technical Report ORG-ID:: dilico.uhamburg_cs.TKRN	3
Technical Report ORG-ID:: dilico.uhamburg_cs.NATS	3
Original clinical investigation	2
Graph	2
Lesson plan	2
Meeting Abstracts	2
Advertisement	2
Textbook	2
audio	2
Bibliography	2
Book chapter	2
Brief report	2
Database	2
Technical Report ORG-ID:: dilico.uhamburg_cs.ASI	2
Technical Report ORG-ID:: dilico.uhamburg_cs.LKI	2
Dissertation	2
Technical Report ORG-ID:: dilico.uhamburg_cs.RZ	1
dataset; activity; field; classroom; computer; lab; assessment; curriculum; instructor; lesson plan; problem set; audio; lecture; module; visualization; reference; report; code; maps; photographs; portal; case study; course; illustrations; video; interactive; webcast; library; dleseTeaching--Aids and devices	1
glossary	1
Supplement Preface	1
Dissertation ORG-ID:: dilico.uhamburg_cs.AGN	1
Example	1
ERCIM-News	1
EC-deliverable LOCLANGUAGE:: Italian	1
EC-deliverable LOCLANGUAGE:: Italian LOCABSTRACT:: Questo documento presenta la versione rivista sia della fase di progettazione . . .	1
EC-deliverable LOCLANGUAGE:: Italian LOCABSTRACT:: Questo rapporto documenta la fase di valutazione del sistema MIAOW rispetto alla . . .	1
Syllabus	1
Technical Innovations	1
Technical advance	1
Short communication	1
Open letter	1
Magister thesis	1
Retraction	1
Non-randomised controlled trial	1
Technical Report ORG-ID:: dilico.uhamburg_cs.WSV	1
Curriculum	1
Technical-Report LOCLANGUAGE:: Dutch	1
Technical-Report LOCLANGUAGE:: Hungarian	1
Technical-Report LOCLANGUAGE:: Hungarian	1
Reports	1

Technical-Report LOCLANGUAGE:: Italian LOCABSTRACT:: Extensible Markup Language (XML)	1
Technical-Report LOCLANGUAGE:: Italian LOCABSTRACT:: Il linguaggio Java . . .	1
Technical-Report LOCLANGUAGE:: Italian LOCABSTRACT:: Il task Applicazioni . . .	1
Technical-Report LOCLANGUAGE:: Italian LOCABSTRACT:: In fisica, molti integrali . . .	1
Cohort study	1
Report ORG-ID:: dilico.uhamburg_cs.WSV	1
case study	1
Technical-Report LOCLANGUAGE:: Italian LOCABSTRACT:: In questa pubblicazione sono . . .	1
images	1
Report ORG-ID:: dilico.uhamburg_cs.TECH	1
Technical-Report LOCLANGUAGE:: Italian LOCABSTRACT:: In questa pubblicazione sono . . .	1
Report	1
Technical-Report LOCLANGUAGE:: Italian LOCABSTRACT:: Internet un mondo virtuale in . . .	1
quiz	1
Technical-Report LOCLANGUAGE:: Italian LOCABSTRACT:: L'introduzione del multicast IP . . .	1
Technical-Report LOCLANGUAGE:: Italian LOCABSTRACT:: Questo lavoro descrive come sono . . .	1
Technical-Report LOCLANGUAGE:: Italian LOCABSTRACT:: Una proposta per la ristrutturazione . . .	1
Publication ORG-ID:: dilico.uhamburg_cs.VSYS	1
Technical-Report LOCLANGUAGE:: Italian LOCABSTRACT:: Vengono richiamate le definizioni e la teoria . . .	1
Technical-Report MSC:: 65H05 Single equations . . .	1
Unit of instruction	1
Technical-Report MSC:: 65U05 Numerical methods in probability and statistics . . .	1
Audio:Audio book	1
Project-Report LOCLANGUAGE:: Italian	1
Technology review	1
Proceeding LOCLANGUAGE:: French	1
Master Thesis ORG-ID:: dilico.uhamburg_cs.VSYS	1
Annotated bibliography	1
Misc	1
Preprint MSC:: 16D60 Simple and semisimple modules, primitive rings and ideals MSC:: 68Q40 Symbolic computation, algebraic computation	1

Table 4 Audience elements in the metadata repository

Element Name	Count
High School (9-12)	12210
Middle School (6-8)	6902
College	5895
Elementary	3343
Research	1714
Late Elem. (3-5)	414
Early College	192
Fifth Grade	106
Sixth Grade	86
Seventh Grade	78
Fourth Grade	77
Graduate	71
Middle school teachers	66
Elementary school teachers	66
Secondary school teachers	64
Eighth Grade	58
First Grade	58
Second Grade	56
Lifelong Learner	55
Early Elem. (PreK-2)	54
Undergraduate	54
Ninth Grade	43
Late College	41
Kindergarten	38
Evaluators	36
Teachers	34
Twelfth Grade	33
Third Grade	31
Pre-Kindergarten	30
University instructors	30
Eleventh Grade	25
Tenth Grade	25
teachers, students	13
Students	10
K-12 teachers	9
K-12; higher ed; teachers; students; public	7
K-12 students; K-12 teachers	5
HighSchool (9-12)	4
General audiences	4
University students and faculty	4
K-12 teachers; K-12 students	4
MiddleSchool (6-8)	3
Researchers	2
K-12 students; K-12 teachers; general	2
K-12; teachers; students	2
K-12 students and teachers; undergraduates;general public	2

K-12 students and teachers; undergraduates; general public	2
K-12; teachers	2
K-12 students and teachers	2
K-12 teachers; university teachers	2
K-12 students	2
K-12; general public; graduate; professional; informal education; undergraduate	2
Middle School(6-8)	2
University teachers and students	2
University faculty and students	2
engineers; scientists; researchers; higher ed	2
Fourth-Year College	2
High School(9-12)	2
Higher education	2
undergraduate; graduate; professional	1
Physics researchers and students	1
Undergraduate; Graduate;K1-K12;researchers;publisher	1
Undergraduates	1
middle school; secondary school; higher education; general public	1
middle school, high school, undergraduate, graduate	1
undergraduates	1
Undergraduates, Faculty	1
undergraduates; faculty	1
undergraduates; graduate students	1
University	1
Medical students and faculty	1
Medical and Health Science students	1
lower division undergraduate	1
teachers	1
librarians; researchers; teachers; higher ed	1
University students	1
University students and teachers	1
students; higher ed; researchers; scientists; engineers	1
teachers; K-12	1
Secondary students and teachers; university faculty and students	1
teachers; k-12; higher ed	1
Teachers;researchers;students	1
Teachers;students	1
Teachers;undergraduate	1
Second-Year College	1
Technical; Professional	1
Technology, Math, and Science Teachers and Educators	1
scientists; general public; high school and university teachers	1
scientists; engineers; teachers; high school students; college students	1
scholars	1
researchers;students;teachers	1
researchers; teachers	1
Third-Year College	1
Undergraduate and graduate level computer science teachers and students	1

undergraduate faculty	1
Professionals; University faculty and students; general public	1
Teachers; general public	1
Primarily undergraduate level with applications for middle school through research level	1
Undergraduate students; Graduate students; Materials Science educators; Industrial community; University Materials Science researchers; Government Materials Science researchers	1
Higher education faculty and students	1
higher ed; researchers; scientists; engineers	1
higher ed	1
High school through undergraduate	1
High school students; higher education; vocational education; general	1
High school and University students; High school and university teachers	1
High School and College students and teachers	1
Graduate or professional; Undergraduate lower division; Undergraduate upper division	1
Grades 7-12; environmental science students; teachers	1
Grade 8-12 teachers; Grade 8-12 students; higher ed	1
grade 6-12; middle school; high school; undergraduate; graduate; professional; researchers	1
General public; teachers; students	1
general public; K-12 students	1
general public	1
General public	1
General audiences; K-12 Teachers and students	1
General	1
First-Year College	1
Faculty, teachers, developers	1
engineers; teachers; students; higher ed	1
Engineering undergraduates	1
Elementary; Middle School; High School; Undergraduate; Graduate; Research; Lifelong learner	1
egineering researchers, teachers, students	1
educators;students; general public	1
Education Levels	1
Computer Science educators and students, primarily at post-secondary institutions	1
College level teachers and students	1
All of Me	1
All levels	1
advanced high school students, college undergraduate and graduate students, researchers	1
Adults; K-12 teachers; K-12 students	1
6-12; middle; junior; high; teachers	1
K-middle school	1
K-12; teachers; students; public	1
K-12; teachers; students; parents; public	1
K-12; teachers; students; parents; librarians	1
K-12; teachers; students; parents; families	1

K-12; teachers; students; parents	1
K-12; teachers; parents; students	1
K-12; teachers; adults	1
K-12; students; teachers	1
K-12; parents; teachers	1
K-12; lower division undergraduate	1
K-12; higher ed; teachers; students; public; researchers; scientists	1
K-12; higher ed; teachers; students; public; parents; researchers	1
K-12; higher ed; librarians; researchers	1
K-12; girls	1
K-12, Higher Ed, Life-long learners. Majority of records are aimed at undergraduate engineering education.	1
Higher education students and faculty	1
K-12	1
K-12 educators; pre-service teachers; teacher educators; parents	1
K-12 Students	1
K-12 students and teachers; undergraduates	1
K-12 students and teachers; undergraduates; graduate students;	1
K-12 students; K-12 teachers; adults	1
K-12 students; K-12 teachers; general public	1
K-12 students; K-12 teachers; scientists	1
K-12 Teachers	1
K-12 teachers and students	1
K-12 Teachers and students; General audiences	1
K-12 teachers and students; parents	1
K-12 teachers and students; University teachers and students	1
K-12 teachers; K-12 students; general	1
K-12 Teachers; K-12 students; general public	1
K-12 teachers; K-12 students; parents	1

Towards Making the NSDL Collection More Accessible Through a Testbed

Peter Shin
San Diego Supercomputer Center, UCSD

The size of the NSDL collection has been steadily increasing. As the number of documents increases, the usability of such a large collection becomes more difficult and locating relevant materials becomes a serious challenge. There are one hundred projects contributing to the NSDL collection. However, a method for evaluating and integrating the results of these projects does not currently exist. This paper is written from the perspective of knowledge discovery using data mining techniques and therefore the meaning of terms used in this paper, such as "classification" comes from the machine learning community. In this paper, we argue for creating a collaborative testbed environment for sharing data, methods, results, and evaluations of such results. First, we describe the data in the NSDL collection, and then motivate the idea of a testbed by identifying the multiple classification dimensions that users (both educators and students) need and by listing the key testbed components that are required.

The NSDL collection is a web-based digital collection of educational resources from multiple websites that encompass science, technology, mathematics, and engineering. It is heterogeneous, distributed, and complex in nature. The San Diego Supercomputer Center (SDSC) has been archiving the NSDL collection and processing the materials to facilitate analysis. The SDSC archive contains copies of the web materials. Each contributing web site is crawled, the retrieved material is deposited into a persistent archive, and all internal links between the retrieved items are converted into handles managed by the persistent archive. In addition, the metadata published through the Open Archives Initiative interface is also retrieved. The metadata describes the set of materials that is located at the URL that is published into the NSDL central repository. The metadata is based on the education extensions to the Dublin Core (DC) Standard. Attributes include grade level, discipline, a short summary, and publication date. A representative sample of the persistent archive has been created that contains the material from about 20,000 URLs. The metadata for the URL is associated with the initial HTML page ("first page").

A first step in characterizing the NSDL documents is identifying the classification dimensions that would be useful to the community of users. Here we present an initial set of classification dimensions and identify associated challenges.

Grade Level Vocabulary and Discipline:
One problem with hand-generated vocabulary and discipline labels is the inconsistency across documents. It is extremely challenging to objectively label documents by hand, and many studies have shown that human

cataloging and indexing are subjective processes which can result in low consistency. An important task for the NSDL community is to evaluate various approaches to validate vocabulary and discipline metadata.

Level of Granularity:
Documents contain complex structures. Often a variety of concepts from multiple disciplines are distributed across the structures. On the other hand, concepts can be limited in scope and may be restricted to a single paragraph within the document. This raises the issue of identifying appropriate document subsections and, more generally, the question of whether document-level micro-classification and depth indexing are needed to meet users' information retrieval needs. A systematic evaluation of multi-resolution classification schemes and of the relationships to relevancy, is needed to address these concerns.

Strand Maps:
A final issue for consideration is the use and utility of other existing schemes for structuring disciplinary knowledge. For example, the American Association for the Advancement of Science (AAAS) has organized and mapped the concepts in every subject according to their grade level and the relationships. Establishing a correspondence between the AAAS classification and the NSDL collection could potentially provide enormous benefits. The ability to employ the AAAS concept maps in the intelligent information retrieval of NSDL documents could greatly improve educators' ability to locate relevant NSDL materials. An investigation into methods for mapping the NSDL collection to the AAAS concept maps is an important task that offers great potential to enhance the utility of the NSDL.

In order to explore these issues, we propose to design and develop a comprehensive testbed for the NSDL community. This testbed will contain 5 key components. These are described below:

1. *Corpus*
   All of the archived data is publicly available as html documents. However, in order to analyze the data, the formatting tags should be removed, and the body text should be stemmed. Our initial study has produced a representative dataset containing documents in text files and documents with stemmed words. Our initial study was conducted on the top pages collected in December, 2003. It contains over 20,000 HTML documents with the associated metadata. This processed collection can serve as a starting point for integrative efforts on characterizing the NSDL dataset.

2. *Classification Dimensions*
   Multiple metrics exist for categorizing the documents. In particular we are interested in classification by grade-level vocabulary, grade-level science concepts, peer review, and scientific discipline concepts. An example of

the latter is the DLESE concept space for earth system science. We suggest an organized activity to apply and evaluate these metrics. We would also like to encourage focused pilot projects with organizations such as DLESE.

3. *Algorithms*
   In order to make the NSDL collection useful, multiple classification algorithms should be applied. Most of the classification algorithms are publicly available, but combining them into single-query methods requires new approaches and perhaps new research activities. Publishing the software tools developed in these new activities will facilitate collaboration and encourage other communities to participate in this investigation.

4. *Hardware*
   In order to support these activities we recommend the development of a physical information technology infrastructure. This infrastructure will include high-performance computers, adequate storage space, and reliable network connections. It is a part of our mission at SDSC to make such resources available to the scientific research community. The high-performance computing resources at SDSC could play a significant role in enabling the collaborative analysis needed by the NSDL community. As part of the core infrastructure team, the NSDL persistent archive is maintained on a disk cache for ready access.

5. *Evaluation with NSDL experts and teachers*
   Validating the results with NSDL end users (science experts, teachers, and students) is a vital component in constructing a useful testbed. The choice of metrics needs to be acceptable to the communities that are assembling the collections. The metrics and classification algorithms need an interface that teachers and educators will be able to interpret and use. The ultimate goal of the testbed is to provide useful services to the users. We recommend that a special interest group be formed that focuses on the design of user tasks and measures for systematic evaluation of the usefulness of diverse automatic classification for information retrieval from the NSDL collection. For example, there is potential to adapt relevance, one of the traditional measures for information retrieval, and then construct usable measures and indicators for evaluation such as on-and-off topic relevancy, cognitive relevance, indicators of interactivity, etc.

Current activities at SDSC

We have initiated a variety of activities in support of the NSDL testbed development. Each month, a crawler collects and archives all the web materials in the NSDL and the materials linked from the NSDL websites to a depth of 10 levels of indirection. We also have built a prototype of a parallel HTML document processing system to summarize the documents into “weighted feature vectors” which various categorization methods will need. In addition, using the “subject”

attribute from the metadata, we have built a classification system with a method called “Support Vector Machines.” Currently, we are engaged in an evaluation of grade-level classification dimension, and the results will be presented at a related session of NSDL Annual Meeting in November, 2004.