Computer-Assisted Assignment of Educational Standards Using Natural Language Processing

Holly Devaul
Digital Learning Sciences, University Corporation for Atmospheric Research, 3300 Mitchell Lane, Boulder, CO 80301. E-mail: devaul@ucar.edu

Anne R. Diekema
Instructional Technology and Learning Sciences, Utah State University, 2830 Old Main Hill, Logan, UT 84322. E-mail: anne.diekema@usu.edu

Jonathan Ostwald
Digital Learning Sciences, University Corporation for Atmospheric Research, 3300 Mitchell Lane, Boulder, CO 80301. E-mail: ostwald@ucar.edu

Educational standards are a central focus of the current educational system in the United States, underpinning educational practice, curriculum design, teacher professional development, and high-stakes testing and assessment. Digital library users have requested that this information be accessible in association with digital learning resources to support teaching and learning as well as accountability requirements. Providing this information is complex because of the variability and number of standards documents in use at the national, state, and local level. This article describes a cataloging tool that aids catalogers in the assignment of standards metadata to digital library resources, using natural language processing techniques. The research explores whether the standards suggestor service would suggest the same standards as a human, whether relevant standards are ranked appropriately in the result set, and whether the relevance of the suggested assignments improve when, in addition to resource content, metadata is included in the query to the cataloging tool. The article also discusses how this service might streamline the cataloging workflow.

Introduction

Since passage of the No Child Left Behind Act of 2001 (2002) in the United States, teachers are required to document that their instruction is aligned with relevant state or national standards. Focus groups conducted by Devaul and Kelly (2003) investigated teachers' interests and needs in such digital library services. The results of these focus groups indicate that teachers want standards information associated with classroom materials and prefer specific information such as detailed national standards or their particular state standards. Digital libraries can address this need by providing access to materials that have been associated with educational standards information and developing services that support the use of these materials in classroom settings.

Of the three main elements that make up a digital library—documents, technology, and work—, Levy & Marshall (1995) identify work as the most essential. Work is defined as the work of digital library users but also the work of librarians to support the users by making the library resources accessible. After all, a collection and supporting technology would be of no value if users could not find what they needed. The research presented here concerns both types of work: (a) the work of teachers to educate students according to certain state and national education standards, and (b) the work of catalogers to associate these standards with educational resources as metadata, thereby making the most relevant resources accessible.

Associating educational standards with learning materials requires that two significant challenges be addressed.

- *Complexity of the standards landscape*. There exists a multitude of educational standards documents, i.e., there is no single vetted set of standards for the United States. There are several national documents in addition to each of the states and many local district levels (Blank & State Education Assessment Center, 1997; National Research Council, 1996; Project 2061, 1993, 2001)
- *Human costs of associating standards with learning resources*. Additionally, the structure and complexity of these

Received December 3, 2009; revised August 16, 2010; accepted September 2, 2010

 • Published online 29 November 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.21437

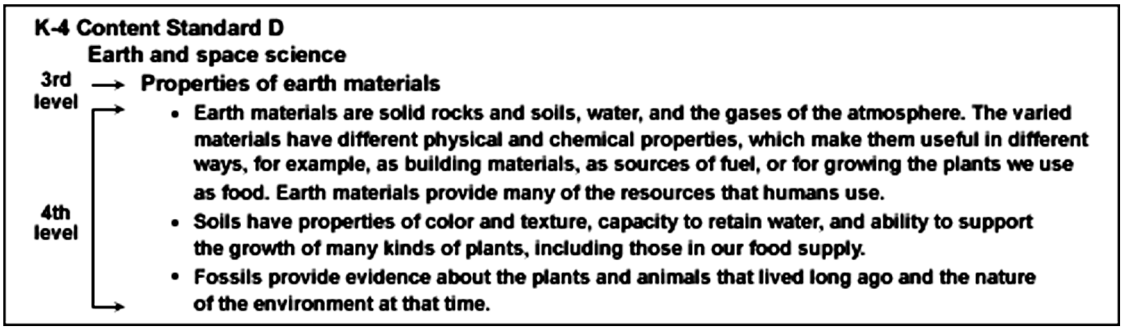

FIG. 1. National Science Education Standards example (National Research Council, 1996).

documents vary widely, presenting the cataloger with a daunting cognitive task in associating relevant standards from the appropriate standards document with the content of a learning resource.

This article describes a cataloging tool, in which a standards suggestor service has been integrated. The service utilizes natural language processing techniques to assess the content of a resource and suggests standards for association with that resource, thereby providing catalogers with a recommended subset of the "standards space" from which to make assignments (Diekema & Chen, 2005; Diekema, Yilmazel, Bailey, Harwell, & Liddy, 2007). In evaluating this tool, we focus on its effectiveness in improving the cataloging workflow that enables access to standards-based educational information in a digital library.

We describe the background and motivation of this work, followed by a detailed explanation of the cataloging tool, the standards suggestor service, and their integration. We then describe an evaluation made regarding the effectiveness of this integration in assisting in the assigning of National Science Education Standards (NSES) to geoscience-focused materials in the context of a digital library cataloging workflow. We conclude with an analysis of results and suggestions for future work.

Background and Motivation

This work is a collaborative effort to address the technical and social challenges of providing efficient assignment of standards information for use in a digital library setting. Funded by the National Science Foundation (NSF) through the National Science Digital Library (NSDL; http://www.nsdl.org), this work supports their mission to provide organized access to high-quality resources and tools that support innovations in teaching and learning at all levels of science, technology, engineering, and mathematics education. The Digital Library for Earth System Education (DLESE; http://www.dlese.org) is the geoscience pathway of the NSDL and has been investigating issues related to the assignment of standards to digital learning objects as well as user needs in accessing this information. The Center for Natural Language Processing (CNLP; http://www.cnlp.org) has applied their natural language processing technology to automatic metadata assignment, with a special focus on computer-assisted assignment of standards metadata.

In response to the critical role that education standards play in the U.S. education system, DLESE initiated the cataloging and discovery of NSES information in association with learning objects in the collection. The NSES were created by the National Academy of Sciences and describe desired student competencies in the different areas of science at different grade levels to ensure a "scientifically literate populace" (National Research Council, 1996).[1] These standards can be represented hierarchically with increasing detail culminating in a statement of learning goals at the fourth level (Figure 1). Cataloging efforts initially focused on the third level of detail of these standards, a topical delineation of a larger domain of science knowledge.

As with any new metadata field with a controlled vocabulary, cataloging standards information requires additional time and skills. Specifically, it requires the cataloger to have knowledge of the standards documents being referenced and, in some cases, the pedagogical approach embodied by the resource. Although there was a cost associated with this additional metadata assignment in terms of time and training, it was accommodated in the DLESE cataloging workflow because the benefit of providing the standards information to users was considered a high priority.

Subsequent focus groups, however, indicated that users preferred more detailed national standards information as well as associated state standards, because the learning goals articulated therein more directly reflect the concepts that drive teaching practice and are incorporated into student testing (Devaul & Kelly, 2003). In response to this finding, a move to both catalog at a more detailed national standards level and provide access to state standards information was undertaken. These changes presented additional technical and cognitive challenges as the complexity of the standards was increased substantially. In moving from the third to the fourth level of the hierarchy of the NSES, the number of standards to display for consideration increases from 87 to over 350, resulting in user interface issues for the cataloging tool, and similarly

[1] http://www.nap.edu/openbook.php?record_id=4962

DOI: 10.1002/asi

impacting the cognitive load on the cataloger in reviewing and selecting appropriate standards for assignment. The increased complexity of the NSES moving from the third to fourth level of the NSES is illustrated in Figure 1, where the fourth level comprises statements that describe learning goals that further expand the third level topical phrase. To address this increased level of detail and complexity, the suggestor service was developed as an approach to aid human catalogers in assigning appropriate educational standards to specific learning materials.

Automatic Metadata Evaluation

Adhering to the adage that correct metadata are a prerequisite for functional metadata, the majority of metadata evaluations concern themselves with the quality or correctness of the metadata (Zhang & Li, 2008) and this study is no exception. Metadata quality might be the focus, but, as pointed out by Hillman (2008), quality criteria should not be viewed independently from the functionality that is required from the metadata. Quality and functionality are, thus, closely tied. According to Park (2009), the most commonly used metadata quality measurement criteria are accuracy, completeness, and consistency, which all contribute to the discoverability of an item.

A commonly used approach for metadata evaluations is to utilize either existing "gold standard" catalog records created by humans or humans to directly evaluate the system's output. Related work on the evaluation of automatic metadata has been carried out by the MetaTest project, an evaluation centered on the comparison of automatically assigned metadata to manually assigned metadata (Liddy et al., 2002). Actual metadata users (science teachers) assessed the quality of the assigned metadata. No statistical difference was found between the quality of automatic and manual metadata. Greenberg (2004) compared the automatic metadata generation (Dublin Core) of two software tools (Klarity and DC.dot) based on a sample of 29 health-related resources. A group of three human catalogers evaluated the systems' output based on whether the metadata would allow the resource to be retrieved by an "intelligent health consumer" or staff of the National Institute of Environmental Health Sciences. Paynter carried out a large-scale automatic metadata extraction evaluation, using the iVia Virtual Library Software (Mitchell, Mooney, & Mason, 2003; Paynter, 2005). iVia software allows users to set parameters as to the type and number of metadata records to be included in the evaluation. These records are then selected from a database that contains a large number of metadata records created by human catalogers. To carry out the evaluation, the metadata extracted by the system was compared with the gold standard metadata. Recent work by Nichols et al. (2009) describes the development of tools to automatically check the quality of metadata records. Reports and visual representations can aid to improve on these records. Unlike the evaluation described in the current article, which only concerns standards metadata, Liddy, Paynter, and Greenberg evaluated multiple metadata fields (Liddy et al., 2002; Nichols et al.; Paynter). Additional single field metadata evaluations concerning Library of Congress Subject Headings (LCSH) have been carried out by Larson (1992) and Frank and Paynter (2004).

Although it is convenient to consider human cataloger output to be the gold standard, it would be naïve to think that perfect agreement between humans is even a possibility (Cooper, 1969; Sievert & Andrews, 1991). According to Olson, Boll, and Aluri (2001), when assigning subject headings, consistency declines when cataloging at a higher level of specificity and a higher level of exhaustivity. In other words, disagreement among human catalogers about what terms to assign to an item increases as these terms get more specific. Similarly, disagreement grows the more thoroughly catalogers are trying to describe an item. In the current study, the cataloging of standards took place at the most detailed level. Also, catalogers assigned large numbers of standards to each educational resource (average of 9, range of 2 to 26). In automatic metadata evaluations, we are evaluating systems on how well they agree with humans even though these humans themselves may not agree completely. This natural variability in human performance needs to be taken into account when examining results of these studies and is considered in the research reported here.

Tools for Cataloging Educational Standards

Our approach to aid catalogers to make standards assignments is embodied in a cataloging tool that integrates the Content Assignment Tool (CAT), developed by CNLP at Syracuse University, within the Digital Collection System (DCS), developed by Digital Learning Sciences at the University Corporation for Atmospheric Research (UCAR). The DCS was initially developed for DLESE, but it is now more widely utilized among diverse digital library partners including the NSDL and its partners.

DCS

The DCS is a Java-based cataloging tool that combines a metadata editor with search, discovery and OAI-PMH[2] dissemination services. The DCS is able to accommodate multiple metadata frameworks that are expressed as XML Schema. The metadata editor interfaces for each framework are automatically generated from the framework's schema and provide a structured view onto the underlying XML metadata records (see Figure 2). For more information on the DCS and its continued development, consult http://ncore.nsdl.org/index.php?menu=services&submenu=services!NCS.[3]

[2]http://www.openarchives.org/pmh/

[3]With ongoing development, the DCS is now referred to as the NSDL Collection System (NCS), as it is as the collection management system of the NSDL. The NCS is available as a metadata management and cataloging tool for projects and CAT is provided as an optional feature.

DOI: 10.1002/asi

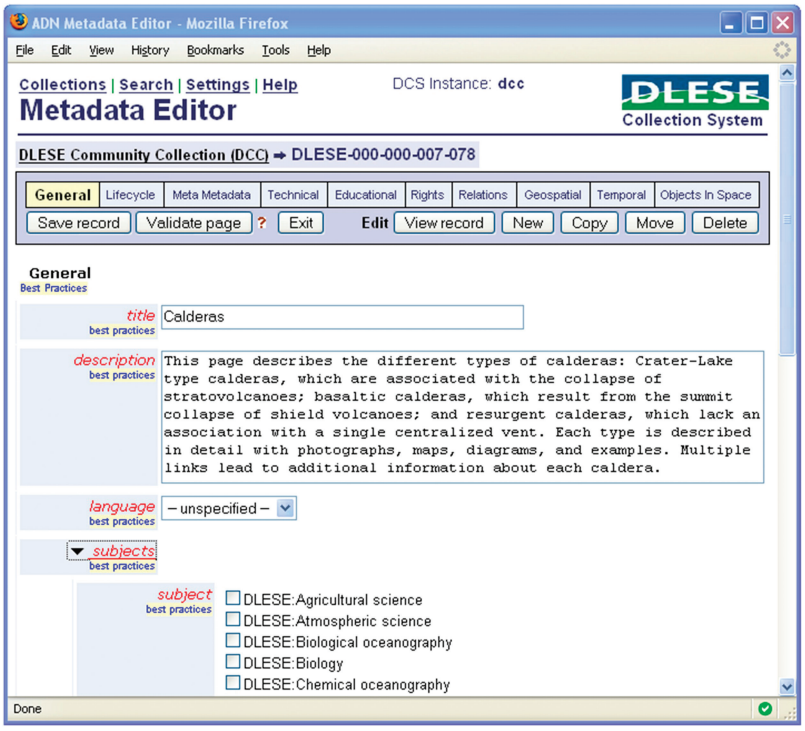

FIG. 2. Metadata editor.

The metadata editor provides many supports for the cataloging process:

- Ensuring well-formed records through schema-based validation of metadata values
- Ensuring that required metadata fields have values
- Providing pick-lists for controlled vocabularies to eliminate typographical errors
- Offering easily accessible best practices information for each field

Educational standards are defined in the metadata frameworks as controlled vocabulary lists. The metadata editor displays these lists as collapsible hierarchies, the leaves of which can be selected to assign a standard to a particular resource. Although the metadata editor can ensure that a cataloger cannot assign an illegal value to a resource, it cannot offer assistance for selecting the *appropriate* standards for a resource. Thus, while the DCS metadata editor addresses many of the syntactical tasks facing a cataloger, when it comes to cataloging standards, more support is needed.

CAT

The CAT was originally developed as a stand-alone application to assist collection providers, catalogers, and teachers in assigning educational standards by assessing the content of a learning resource and making suggestions as to which standards may be most relevant. These standards are then manually reviewed and selected for association with that resource (Diekema & Chen, 2005; Diekema et al., 2007). CAT utilizes CNLP's in-house natural language processing software TextTagger (Yilmazel et al., 2007) to process both the standards and the educational resource to facilitate more accurate matching between the resource and the standards. TextTagger is a rule-based information extraction system developed at the Center for Natural Language Processing at Syracuse University. TextTagger processes (unstructured) electronic text in eight sequential phases to extract lexical, syntactic, and semantic information. During each phase, special processing is carried out to feed into the next phase. The phases are described below and an example from a 2009 Tennessee English language arts benchmark processed by TextTagger can be found in Table 1.

The tokenization phase splits the text into individual words or tokens. Groups of tokens are joined into sentences during the sentence detection phase. After TextTagger has marked the sentences, the components are tagged with part-of-speech information. After that the components of the sentence are turned into their lemmatized form (e.g., strategies would become strategy). During the next phase, the system identifies noncompositional phrases (e.g., sentence structure), because the individual words making up these phrases have different meanings from the phrase as a whole. Next, the system brackets compositional phrases and recognizes temporal concepts, numeric concepts, named entity phrases, and common nouns. Named entities are categorized (people, materials, substances) in the next phase. Finally, the system extracts (named) entities, relations between these entities, events, and relationships between these events. CAT uses only information from the earlier phases of processing (tokenization, sentence detection, part-of-speech tagging, lemmatization, phrase detection, named entity categorization).

DOI: 10.1002/asi

TABLE 1. TextTagger output.

Benchmark	Demonstrate control of standard English usage, mechanics, spelling, and sentence structure
Tagged benchmark	Demonstrate|VB <CN> control|NN </CN> of|IN <CN> <NP cat="unknown"> Standard|NP English|NP </NP> usage|NN </CN> ,|, <CN> mechanics|NNS </CN> ,|, <CN> spelling|NN </CN> ,|, and|CC <CN> sentence_structure|NN </CN> .|.
Terms for matching	Demonstrate; control; standard English; standard English usage; mechanics; spelling; sentence structure

The state and national standard come from the Achievement Standards Network (ASN), which is a national repository of machine-readable educational standards represented in the Resource Description Framework (RDF) language (Sutton & Golder, 2008).

CAT incorporates continuous assignment quality improvement through a custom-built, instance-based machine-learning algorithm (formerly known as "More Like This"). This type of learning algorithm postpones processing until, for example, a new lesson plan needs to be processed. Each final assignment of standards to a resource is stored in a relational database. This information is then utilized by the system via an algorithm similar to k-Nearest Neighbor (kNN), which uses Euclidian distance to find similar lesson plans and their assignments to inform and improve future assignments (Yang, 1999). This learning can take place at the single user level, or at the organizational level where all assignments from multiple users in an organization are aggregated.

Although learning was disabled during the course of this study, an initial round of learning, based on 88 resources, was incorporated into the algorithm. Human catalogers from DLESE supplied the training data.

A retrieval test was done to determine the contribution of the NLP technology to the retrieval of the standards. We compared a basic no frills CAT, very similar to a standard tf*idf retrieval algorithm, with the regular CAT, which contains natural language processing capabilities. The test used the same 29 queries that were used in the rest of this study. The differences between the standard retrieval algorithm and CAT were small. This is not surprising, given the limited number of test queries at our disposal. We ran the trec-eval program (http://trec.nist.gov/trec_eval/) on the retrieval results to calculate the different evaluation metrics. The two algorithms did not show any detectable differences in the ratio of relevant standards to all standards returned (precision). Slight differences between algorithms showed up in the interpolated recall-precision measures. Although both algorithms retrieved exactly the same number of relevant standards the no frills CAT algorithm ranked relevant standards higher in the result set.

In a typical usage scenario, a cataloger wishes to assign standards to a particular educational resource. The user specifies the standards they want to assign (e.g., New York Math or National Science Education Standards) as well as the grade level and location (URL, file path) of the resource they are cataloging. The user then selects the "Suggest Standards" button, and CAT responds by presenting a list of suggested standards, ranked by their relevance to the content of the specified resource. The user reviews the list of suggested standards and selects those that he or she wants to assign. If the user wishes to assign additional standards beyond those suggested by the system, a navigable standards tree is available to browse the standards text and make additional selections. All selected standards are then associated with the resource URL in the CAT database.

Although the stand-alone CAT tool provides a mechanism to access and assign educational standards, it is not a cataloging tool that generates metadata records connected to a library discovery system. Making this connection to reduce steps in the cataloging process and to bring the standards information into library operations was the next step in the collaboration.

DCS-CAT Integration. To combine the strengths and address the limitations of the respective systems, the standards assignment capabilities of CAT were integrated into the DCS cataloging tool to create a seamless workflow (see Figure 3).

To facilitate this integration, the stand-alone version of CAT was re-implemented to separate the user interface from the underlying standards suggestion and assignment functionality. To CAT users, this separation is transparent. However, in this implementation the CAT front-end communicates with a standards suggestion REST Web Service API (CAT Service) that is responsible for making standards suggestions for a given resource, and for storing standards assignments in a database (Rodriguez, 2008). The CAT Service is accessible not only to the CAT front end, but it is also available to the NSDL and its affiliated projects and other Web-based systems.

The CAT Service functionality that is most relevant to this article is the *getSuggestions* function, which returns an ordered list of standards that are most relevant to a provided resource URL. In addition to the resource URL, a *getSuggestions* request may contain several other pieces of information to influence the suggestions returned by the CAT Service. For example, providing a *GradeRange* will cause the CAT Service to return only standards within that range.

The DCS-CAT tool uses the *getSuggestions* function to obtain a rank-ordered set of standards that are relevant to the resource currently being cataloged. The resource URL is extracted from the metadata record, and sent in a *getSuggestions* request to the CAT Service, which returns a list of suggested standards. The suggested standards are then displayed in the metadata editor interface, where they can be

DOI: 10.1002/asi

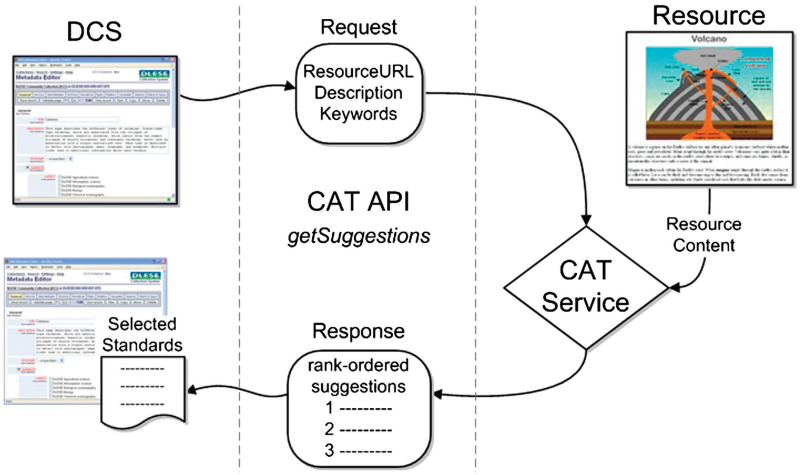

FIG. 3. DCS-CAT integration.

Suggestion Criteria		
Criteria	**Enabled**	**Value**
grade ranges [reset]	☑	start grade: Middle school (6-8) — end grade: High school (9-12)
keywords [reset]	☑	troposphere
description	☑	Using current value of description field
resource Url		http://www.ucar.edu/learn/1_7_2_28t.htm

FIG. 4. DCS-CAT metadata specifications.

accepted or rejected by the cataloger before being saved to the metadata record.

In addition to the resource URL, the metadata record can supply other information to the CAT Service in the *getSuggestions* request, including a detailed resource description, keywords, and grade range. Although the grade range metadata constrains the suggested standards, as described above, the description and keywords extracted from the metadata record provide the natural language processing algorithm, underpinning the CAT Service, with additional input when determining relevant standards. This ability to augment the textual information available to the CAT Service allows suggestions to be obtained for resources with little to no text but for which detailed metadata exist. Figure 4 shows the DCS-CAT interface for specifying what metadata, if any, should be provided to the CAT Service. By default, the criteria shown in the figure are enabled, and initialized with information from the metadata record. The cataloger can change these defaults at any time and submit a new *getSuggestions* request for updated Suggestion Criteria.

The metadata editor interface for viewing and displaying standards was modified to address the following challenges: how to keep track of selected standards, how to select standards that were not suggested, and how to differentiate suggested standards from those not suggested by the service. The mechanism for viewing and assigning standards in the DCS-CAT interface was developed collaboratively by CNLP and DLESE. There are three ways to view standards:

- **Suggested Standards**—a list of suggested standards (those returned by the most recent *getSuggestions* request), ordered by relevance.
- **Selected Standards**—a list of only those standards that have been selected (assigned) by the user. This list may contain suggested standards as well as additional standards selected by the user.
- **Browse**—the entire set of available standards is displayed in a collapsible hierarchy. In this view, the user can select or deselect any standard, whether suggested or not.

Visual cues were developed to assist users in tracking the workflow when assigning standards (Figure 5). The cues indicate which standards were suggested by the CAT Service (shaded in grey), which from this group were assigned (check boxes selected), and which standards were selected independently by using the Browse interface (no shading but check box selected).

Evaluation

A study was designed to assess the use of the DCS-CAT tool as a workflow support to aid in assigning educational

DOI: 10.1002/asi

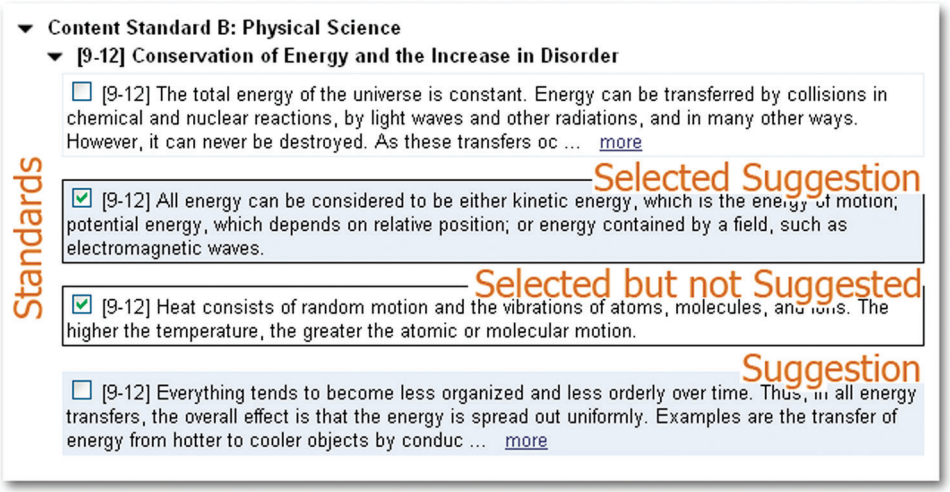

FIG. 5. Visual cues for assigning standards.

TABLE 2. Example grades 5–8 National Science Education Standards from each subgroup.

Inquiry	Standard A Science as Inquiry: Understanding about scientific inquiry: Technology used to gather data enhances accuracy and allows scientists to analyze and quantify results of investigations.
Subject	Standard D Earth and Space Science: Structure of the earth system: Water is a solvent. As it passes through the water cycle it dissolves minerals and gases and carries them to the oceans.
Applied	Standard F Science in Personal and Social Perspectives: Natural hazards: Internal and external processes of the earth system cause natural hazards, events that change or destroy human and wildlife habitats, damage property, and harm or kill humans.

standards to digital learning resources. To do this, we compared assignments made by humans with the suggestions provided by the CAT Service. We also looked at the effect of including additional metadata in the query to the CAT Service and examined the tool's efficacy in suggesting subsets of NSES.

Two catalogers, both trained at DLESE and experienced in standards assignment, were tasked with independently assigning NSES at the detailed fourth level of hierarchy to each resource. These data could then be compared with one another for inter-rater reliability purposes, as well as with the suggestions provided by the CAT Service.

Three research questions were addressed:

- Does the CAT Service return standards that a human would assign? A measure of this is simply the percentage of the human-assigned standards that were also suggested by the tool.
- Do the standards that human catalogers have assigned occur near the top of the ranked list of 10 CAT Service suggestions? A measure of this is the average precision of the CAT service suggestions, which can be computed for each rank and for different data sets.
- Does the inclusion of metadata improve standards assignment via the CAT Service? A measure of this is the comparison between assignments with and without the use of metadata against human assignments.

The Resources

A group of 35 educational resources were selected to span topics in earth and space science. This number later declined to 29 because of issues with accessibility and changing content. Resource selection criteria included an emphasis on text-based content for machine readability, a range of grade levels distributed across K-12, and a range of earth system science topics (weather and water, environment, geology, etc.). These criteria were identical to those used to select the 88 resources used as training data in earlier development work. There was no overlap between the training and the testing resource sets.

The Standards

The NSES are hierarchical in structure and include over 350 individual standards at the fourth level of hierarchy (National Research Council, 1996). They are classified by grade level and content area (which is identified by a letter). These content areas reflect the broad spectrum that comprises science, including the process and application of science. For the purpose of our analysis, we identified three subgroups of the lettered standards (see Table 2 for examples):

- *inquiry* (nature of science; Standard A)
- *subject* (life, physical, earth and space science; Standards B, C, and D)
- *applied* (science and technology, personal and social perspectives, and the history and nature of science; Standards E, F, and G).

Although this evaluation focused on the assignment of NSES standards as a whole, we also explored the performance of the CAT service within the three subgroups of NSES. The premise of this was that interpreting resource content

DOI: 10.1002/asi

in the different areas involves different subjective cognitive tasks that may be reflected in both human inter-rater reliability and CAT's ability to assess a match between resource content and standards.

Assigning Standards to Resources

In the DLESE library, the cataloging protocol for assigning standards to resources is described as follows:

> The association of a standard with a resource signifies that the content of the resource supports the student learning and attainment of the specific ability noted. This can be through many different mechanisms and resource types, including access to background and text-based material as well as inquiry-based activities. Some standards are general in nature, some more specific. The resource need not address the entire scope of the standard for the association to be made, and some resources may not map to any standards at all.

This protocol was adopted by DLESE subsequent to user focus groups across a range of library users and has been the operating cataloging protocol for several years (Devaul & Kelly, 2003). Both catalogers in this study were well-versed in standards cataloging and had jointly assigned the standards to the set of training data introduced to CAT while in development.

Each resource selected for this study was an existing DLESE library resource already described by a metadata record; hence, descriptive metadata such as resource description, subject, and grade level were already available, and the catalogers' task was solely to assign standards at the detailed fourth level. A set of metadata records was created for each cataloger, and the cataloger was allowed to consult the record when making standards assignments, in addition to the resource content. The two catalogers worked independently of one another in assigning standards to each resource.

Duplicate sets of the 29 resource records were also used to query the CAT Service for suggested NSES standards at the fourth level. One set submitted the URL only, thus limiting the NLP algorithm to available resource content (CATnoMetadata), while another included metadata, i.e., the textual description, keywords, and grade level from the record (CATwithMetadata). The 10 ranked suggested standards were retained in order in each record for further analysis.

Two additional sets of records were then generated from the two human-assigned sets: *Common*, which comprises the intersection (agreement) between the two human catalogers in assigned standards to each resource, and *Composite*, which comprises the union of all standards assigned by the human catalogers to each resource.

Results

In the Results section, we report human inter-rater reliability (overall and by NSES subgroup) and then examine both the recall and the precision of the CAT Service compared with human assignments of both the Common and the Composite records, with and without the inclusion of metadata. Four combinations of comparisons were analyzed and trends are reported.

TABLE 3. Inter-rater reliability of two human catalogers.

Overall	Inquiry standards (A)	Subject standards (B, C, D)	Applied standards (E, F, G)
32%	*29.7%*	*40.4%*	*18%*

Note. N = 29 resources, NSES at the 4th level. Number of standards assigned by each ranged from 2 to 26 for a particular resource.

Human Inter-Rater Reliability

Inter-rater reliability expresses the degree to which the two humans agreed on the standards assigned to the resources. We calculated this for all of the NSES and by subgroup of standards, i.e., inquiry, subject, and applied (Table 3). This was accomplished by comparing the specific standards assigned to each resource by each cataloger, and then computing the percentage agreement for the 29 resource records as a whole.

Overall, the catalogers agreed on 32% of the standards selected, varying between 18% and 40% for the different subgroups of standards. This level of agreement appears quite low until one takes into account that for each resource, catalogers make a selection of over 350 potential standards. Also, there is no accepted practical limit as to how many standards should be assigned to a resource, and this judgment can vary between catalogers. The subject standards provided the greatest agreement at 40%. The catalogers agreed on 30% of the inquiry standards and only 18% of the applied standards. Our interpretation of this is that the subject standards typically articulate clearly defined basic science concepts, not requiring significant cognitive interpretation when compared with the content of a resource. Likewise, the subject-focused science concepts covered in a resource also tend to be textually explicit, and they are easily mentally mapped to a subject standard during the cataloging process. The concepts embodied by the inquiry and applied standards are often implicit in a resource, gleaned from analyzing the pedagogical approach of the lesson or the context of the science content. Identifying these components of a resource requires deeper engagement with the resource and perhaps more subjective interpretation.

The inter-rater reliability scores between the two human catalogers illustrate the difficulty in assigning standards to educational resources. Assignment of standards is an extremely subjective task, and different collection groups and educational organizations have different approaches and cataloging guidelines. Even with significant calibration efforts, especially when the standards being assigned are numerous and wide-ranging in scope, as is the case with NSES, strong inter-rater reliability can be difficult to achieve.

Does the CAT Service Return Standards That a Human Would Choose for Assignment?

To answer the first research question, we simply tallied the number of suggestions per resource (N = 29) that were shared

DOI: 10.1002/asi

TABLE 4. Percentage of human-assigned standards also suggested by the tool.

	With metadata	No metadata
Common	51%	46%
Composite	31%	28%

by the CAT Service and (a) both human catalogers (Common) or (b) either human cataloger (Composite) to represent recall. We ignored the difference in the number of standards assigned as well as the rank order of the assigned standards offered by the CAT Service. By default, the CAT Service suggests 10 standards in rank order, while humans were not instructed to follow any minimum or maximum number of assignments or rank their assignments in order of importance or relevance. Humans assigned, on average, nine standards per resource (range of 2 to 26). It is important to note that we did not examine the standards, suggested by CAT, that were not among the standards assigned by human catalogers. It is conceivable that at least some of these suggestions are, in fact, correct but were overlooked by the human catalogers who assigned their standards independent of any computer assistance.

Results. CAT Service suggestions include 28–51% of the standards humans assigned, performing better when metadata is included in the request (see Table 4). From a cataloging workflow perspective, this offers a time savings in the initial assignment of the standards by presenting the cataloger with a smaller subset of standards to choose from, which includes a substantial number of the preferred standards.

Agreement with the Common set of records was greater than the Composite set, and including metadata in the query to the CAT Service improved the returns by 3–5%. These data indicate that in its simplest form as a workflow support, a cataloger can expect to be presented approximately 30–50% of the necessary standards in the initial return from the CAT Service. We did not test CAT's "More Like This" function, which iteratively refines the suggestions based on a cataloger's acceptance of an initial subset, but this could offer additional time savings. Our conclusion at this step in the analysis is that although the CAT Service is not at the point where auto-assignment could be recommended, this level of workflow support offers considerable time savings in the cataloging process.

Do the Standards Human Catalogers Have Chosen Occur Near the Top of the Ranked List of 10 CAT Service Suggestions?

To answer the second research question, we carried out an average precision analysis of the human-assigned standards with respect to their occurrence and placement in the ranked list of 10 suggestions. The assumption is that it would further improve the cataloging process if the appropriate standards appeared preferentially at the top of the list, as is the intent

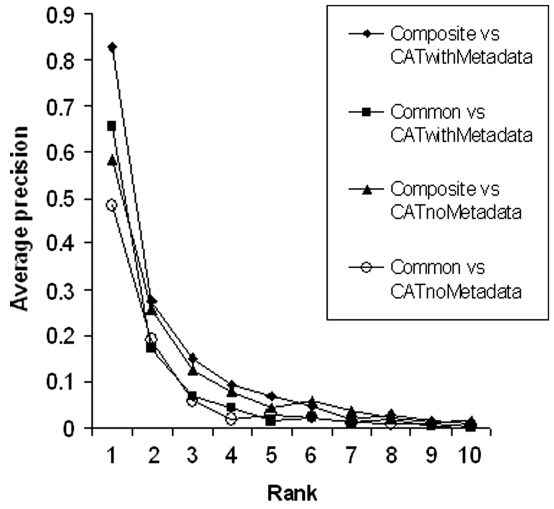

FIG. 6. Average precision of suggested standards for different sets.

of the ranking. We looked at the occurrence of the human-assigned standards rank by rank over all 29 resource records for each of the four record set-pairs (Common and Composite compared with both CATwithMetadata and CATnoMetadata) and for each subgroup of standards (inquiry, subject, and applied science). Average precision was calculated for each rank by tallying the number of human-assigned standards that occurred in each rank as well as those above it. That is, for rank 1, we tallied the number of "hits" that occurred in the number 1 rank, while for rank 3, we tallied the number of hits that occurred in ranks 1, 2, and 3 and averaged over all 3 ranks combined.

Results. Average precision varies among the four record set-pairs, but drops off quickly past rank 1 or 2 for all. The Composite record set (the union of all human assignments) with metadata included in the query produced the highest average precision for rank 1 at 0.83 (Figure 6). The data indicate that CAT's top-ranked suggestion has a good chance of being one that a human would assign; but, beyond that, the standards that a human would assign are distributed throughout the remaining nine ranks and not necessarily near the top.

In all record set-pairs (Common, Composite, with or without metadata) subject standards were suggested at a higher average precision for rank 1 than applied or inquiry standards. This is consistent with the human inter-rater reliability data, in which humans were in agreement on the assignment of subject standards more so than for applied or inquiry standards. In three of the four record set-pairs, inquiry standards were suggested more reliably than applied standards, but both were quite low so as to be inconsequential. This suggests that the natural language processing techniques operate well with explicit science concepts but may be challenged to interpret pedagogical and contextualized aspects of resource content and map them to the standards. We additionally speculate that the resources themselves may not offer the same level of

DOI: 10.1002/asi

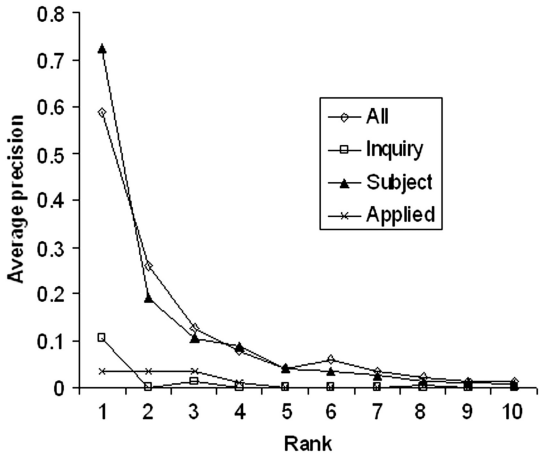

FIG. 7. Average precision [composite record set] without metadata for all NSES and by subset.

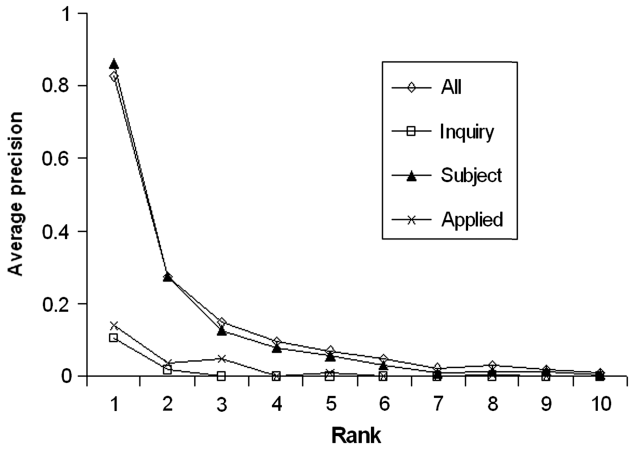

FIG. 8. Average precision [composite record set] with metadata for all NSES and by subset.

explicit text in these areas for NLP algorithms to incorporate in the analysis.

Does the Inclusion of Metadata Improve Standards Assignment via the CAT Service?

To answer the third research question, we compared the CAT Service suggestions with and without the inclusion of descriptive metadata. Including metadata constrains the search for standards to specific grade levels and provides text in addition to resource content that can be used by the NLP algorithms that produce the suggested list of standards.

Results. Including metadata in the query to the CAT Service resulted in a higher initial precision at rank 1 compared to excluding metadata, for both the Common and the Composite record sets. Figure 6 depicts data for all NSES standards combined for each of the record set-pairs. Average precision graphs for the Composite record set without metadata (Figure 7) and with metadata (Figure 8) also illustrate the differences among the three subgroups of standards.

In examining the difference, we see that the average precision scores of rank 1 increase by 0 to 24 percentage points when metadata is included in the query to the CAT Service (Table 5). Based on these findings, we see potential for CAT to be able to suggest standards for nontextual materials such as images based on their descriptions and existing metadata.

Interestingly, the suggestions of inquiry standards remained unaffected by the use of metadata. The metadata used in this study included resource description, grade level, and keyword. None of these fields are explicitly designed to capture information about the scientific process or pedagogy represented in the resource, and, hence, these fields did not provide sufficient additional information to influence the inquiry standards suggested by the CAT Service.

TABLE 5. Increase in the average precision of the rank 1 suggested standard as a result of including metadata in the query to the CAT service.

Standard type	Percentage point increase for Common data set	Percentage point increase for Composite data set
All NSES	18%	24%
Subject standards	10%	14%
Applied standards	N/A; both were zero	10%
Inquiry standards	Unaffected	Unaffected

Note. NSES = National Science Education Standards.

Conclusions and Future Research

The CAT Service offers catalogers an advantage in the complex task of assigning standards to digital learning materials. As a suggestor tool, it currently can provide 30–50% of the standards a human would assign, presented in short lists of 10 for efficient review. This can reduce the search time necessary to identify standards in large, complex standards lists. A navigable standards tree, present in both the stand-alone CAT tool and the DCS-CAT, organizes the standards in hierarchical format, further assisting catalogers in assigning standards.

The DCS-CAT metadata editor offers numerous supports for syntactical tasks to minimize errors in vocabulary and record structure, and when integrated with the CAT Service, it provides support for the cognitive task of assigning educational standards. The interface, enhanced with visual cues, additionally allows catalogers to track workflow and access the long list of NSES standards at a relevant entry point. The CAT Service suggests subject standards with greater precision, and including metadata in the request to CAT further improves the performance of the suggestor service. Inquiry and applied standards may be more difficult to suggest using this technology, compared with subject standards. This is because assigning inquiry and applied standards to a resource may not be textually explicit (i.e. the standard and the resource do not have terms in common) and may

DOI: 10.1002/asi

require additional (subjective) human interpretation to make the assignment.

The next step in the development of this technology is to more fully exercise the machine-learning capability of CAT by introducing additional vetted data into the system. This should increase the percentage of appropriate standards returned in the tool and improve the average precision of the suggestions. This would further streamline the cataloging process by reducing search and selection time. If sufficiently robust, then these refinements could eventually support the automatic assignment of standards to digital resources.

Acknowledgments

This article is based upon work supported by the National Science Foundation under Grant No. 0435339. We acknowledge Tamara Sumner, Sebastian de la Chica, and Elizabeth Liddy, for their guidance and contributions to study design and analysis, and Niranjan Balasubramanian, Jennifer Bailey, Steve Rowe and Wenbo Zhang for their software engineering efforts.

References

Blank, R.K., & State Education Assessment Center. (1997). Mathematics and science content standards and curriculum frameworks: States progress on development and implementation: 1997. Washington, DC: Council of Chief State School Officers.

Cooper, W.S. (1969). Is interindexer consistency a hobgoblin? American Documentation, 20(3), 268–278.

Devaul, H., & Kelly, K. (2003, October). Searching by educational standards in DLESE: What does it mean and what do users want? Poster presented at the NSDL Annual Meeting, National Science Digital Library Annual Meeting. Washington, DC.

Diekema, A.R., & Chen, J. (2005). Experimenting with the automatic assignment of educational standards to digital library content. Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital libraries. New York: ACM Press.

Diekema, A.R., Yilmazel, O., Bailey, J., Harwell, S.C., & Liddy, E.D. (2007). Standards alignment for metadata assignment. Proceedings of the Seventh ACM/IEEE-CS Joint Conference on Digital libraries (pp. 398–399). New York: ACM Press.

Frank, E., & Paynter, G.W. (2004). Predicting Library of Congress classifications from Library of Congress subject headings. Journal of the American Society for Information Science and Technology, 55(3), 214–227.

Greenberg, J. (2004). Metadata extraction and harvesting—A comparison of two automatic metadata generation applications. Journal of Internet Cataloging, 6(4), 59–82.

Hillman, D.I. (2008). Metadata quality: From evaluation to augmentation. Cataloging & Classification Quarterly, 46(1), 65–80.

Larson, R.R. (1992). Experiments in automatic Library of Congress classification. Journal of the American Society for Information Science, 43(2), 130–148.

Levy, D.M., & Marshall, C.C. (1995). Going digital: A look at assumptions underlying digital libraries. Communications of the ACM, 38(4), 77–84.

Liddy, E.D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N.E., et al. (2002). Automatic metadata generation and evaluation. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press.

Mitchell, S., Mooney, M., & Mason, J. (2003). iVia open source virtual library system {computer file}. D-Lib Magazine, 9(1).

National Research Council. (1996). National science education standards: Observe, interact, change, learn. Washington, DC: National Academy Press.

Nichols, D.M., Paynter, G.W., Chan, C.-H., Bainbridge, D., McKay, D., & Twidale, M.B., et al. (2009). Experiences in deploying metadata analysis tools for institutional repositories. Cataloging & Classification Quarterly, 47(3), 229–248.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 107th Cong. §1001 (2002).

Olson, H.A., Boll, J.J., & Aluri, R. (2001). Subject analysis in online catalogs. Englewood, CO: Libraries Unlimited.

Park, J.-r. (2009). Metadata quality in digital repositories: A survey of the current state of the art. Cataloging & Classification Quarterly, 47(3), 213–228.

Paynter, G.W. (2005). Developing practical automatic metadata assignment and evaluation tools for internet resources. ACM/IEEE-CS Joint conference on Digital Libraries (pp. 291–300). New York: ACM Press.

Project 2061. (1993). Benchmarks for science literacy. New York: Oxford University Press.

Project 2061. (2001). Atlas of science literacy. Washington, DC: American Association for the Advancement of Science/National Science Teachers Association.

Rodriguez, A. (2008). RESTful Web services: The basics: IBM Corporation. Retrieved October 13, 2010, from http://www.ibm.com/developerworks/webservices/library/ws-restful/

Sievert, M.C., & Andrews, M.J. (1991). Indexing consistency in information science abstracts. Journal of the American Society for Information Science, 42(1), 1–6.

Sutton, S.A., & Golder, D. (2008). Achievement standards network (ASN): An application profile for mapping K-12 educational resources to achievement standards. In Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications (pp. 69–79). Berlin, Germany: Humboldt University. Humboldt University, Berlin, Germany.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Information Retrieval, 1, 69–90.

Yilmazel, O., Ingersoll, G., & Liddy, E.D. (2007). Finding questions to your answers. In Proceedings of the IEEE 23rd International Conference on Data Engineering (pp. 755–759). Washington, DC: IEEE.

Zhang, Y., & Li, Y. (2008). A user-centered functional metadata evaluation of moving image collections. Journal of the American Society for Information Science and Technology,59(8), 1331–1346.

DOI: 10.1002/asi