Automatically Characterizing Resource Quality for Educational Digital Libraries

Steven Bethard
University of Colorado
594 UCB
Boulder, CO, USA
steven.bethard@
colorado.edu

Philipp Wetzler
University of Colorado
594 UCB
Boulder, CO, USA
philipp.wetzler@
colorado.edu

Kirsten Butcher
University of Utah
1705 Campus Center Dr
Salt Lake City, UT, USA
kirsten.butcher@
utah.edu

James H. Martin
University of Colorado
430 UCB
Boulder, CO, USA
james.martin@
colorado.edu

Tamara Sumner
University of Colorado
594 UCB
Boulder, CO, USA
tamara.sumner@
colorado.edu

ABSTRACT

With the rise of community-generated web content, the need for automatic characterization of resource quality has grown, particularly in the realm of educational digital libraries. We demonstrate how identifying concrete factors of quality for web-based educational resources can make machine learning approaches to automating quality characterization tractable. Using data from several previous studies of quality, we gathered a set of key *dimensions* and *indicators* of quality that were commonly identified by educators. We then performed a mixed-method study of digital library curation experts, showing that our characterization of quality captured the subjective processes used by the experts when assessing resource quality for classroom use. Using key indicators of quality selected from a statistical analysis of our expert study data, we developed a set of annotation guidelines and annotated a corpus of 1000 digital resources for the presence or absence of these key quality indicators. Agreement among annotators was high, and initial machine learning models trained from this corpus were able to identify some indicators of quality with as much as an 18% improvement over the baseline.

Categories and Subject Descriptors

H.3.6 [**Information Systems**]: Library Automation; H.3.7 [**Information Systems**]: Digital Libraries—*Standards, User issues*; I.2.7 [**Computing Methodologies**]: Natural Language Processing—*Text analysis*; I.5.4 [**Computing Methodologies**]: Applications—*Text processing*

General Terms

Human Factors, Algorithms

Keywords

Quality, learning resource, machine learning, educational digital library

1. INTRODUCTION

One goal of educational digital libraries is to develop and curate collections of "high quality" web-based resources useful for teaching and learning across a wide range of grade levels and educational settings. These resources include textual materials (background readings, references), interactive and visual materials (maps, animations, simulations), classroom and laboratory activities, and scientific data. They are created by individuals and institutions across a range of academic, government and non-profit sectors, and submitted to libraries for further dissemination into educational settings. The National Science Digital Library (NSDL), the Digital Library for Earth System Education (DLESE), and MERLOT are prominent examples of such libraries.

Vetting the quality of submitted resources is often a critical issue for educational libraries, but one that is very challenging in practice to carry out reliably and efficiently at the necessary scale. For instance, MERLOT has amongst the most comprehensive peer-review processes for vetting resources, consisting of 15 editorial boards. Yet recently published data highlight the challenge MERLOT is facing in scaling up its peer-review process to keep up with the rate of contributions. Carey and Hanley report that the ratio of submitted to reviewed contributions is approximately eight to one [5]. Interest in scalable methods for evaluating quality have grown in importance with the rise of user and community generated content. Wikipedia, which depends entirely on user contributions, has spawned numerous research efforts exploring algorithmic approaches for automatically assessing article quality [1, 29, 4].

In addition to scalability, another challenge educational libraries face when assessing resource quality is the matter of perspective: the definition of quality is contextual. It depends on the alignment between the user constituency being served, the educational setting where deployed, and the intended purpose of the resource. Acknowledging that there are multiple perspectives on quality, *Connexions* allows third parties to create *lenses*, which provide different views onto collections. For instance, a professional society has created a lens to view resources that have been vetted by their own peer review processes. At the same time, others have created a different lens based on an algorithmic assessment of resource popularity.

Our research is exploring algorithmic approaches for characterizing resource quality to address the dual challenges of scalability and multiplicity of perspective. In broad terms, we are developing methodologies for identifying *dimensions* of quality associated with specific educational purposes and developing algorithms capable of characterizing resources according to these dimensions. Producing an overall quality assessment is not the main goal of this effort; rather our aim is to create a rich, multi-dimensional characterization of different aspects of a resource's quality. These characteristics can then be selectively mixed and matched, by software developers or end-users, to support the next generation of repository management applications and cognitive tools for library users.

In this article, we report on our efforts to date to automatically characterize the value of educational resources for use in formal classroom settings, specifically secondary science education. We conducted a two part study — a "meta-analysis" followed by a study of human experts — to identify the characteristics most commonly used by science educators and science librarians when identifying resources for use in the classroom. Examples of some of the characteristics identified include age appropriateness, inclusion of hands-on activities, and organization around clear learning goals. We then developed prototype algorithms based on supervised machine learning approaches to automatically recognize seven of the identified characteristics. We discuss the strengths and limitations of our work to date, and review future plans for extending this research.

2. RELATED WORK

Two major strands of prior work inform our approach. First, there is a body of work exploring the way experts and users assess the quality of digital resources and web sites. These studies seek to identify both the major factors influencing human judgments and the lower level features of resources that people attend to when assessing quality. A second body of work considers the applicability of machine learning algorithms for this type of problem. Here, there are several inter-related issues, including the consistency (or not) of humans when making potentially subjective assessments and the different features that researchers have chosen to focus on when training machine learning algorithms.

Inherent in our approach is the assumption that when asked to assess the quality of a digital resource or web site, people draw on a suite of often implicit criteria to guide their judgment. Numerous studies have documented this phenomenon. Fogg and colleagues studied how people assess "web site credibility". Using an online survey of over 1400 participants, they identified criteria such as ease-of-use, expertise (of the site), and trustworthiness as being important factors influencing human judgment [18]. Rieh studied how people judge "cognitive authority" while searching for information on the web [25]. This study identified a slightly different set of factors as important for this specific judgment task, including accuracy, currency, trustworthiness, scholarliness and authoritativeness. Sumner and colleagues conducted focus groups to identify criteria to guide collections policy decisions in educational digital libraries serving both formal and informal educational audiences [27]. The criteria identified here for assessing resource quality included scientific accuracy, lack of bias, and good pedagogical design.

This body of work highlights two important issues for our approach. First, even this small sample of reviewed literature highlights the multiplicity of potential perspectives for quality, depending upon the use and audience of materials. Credibility is important when selecting an online bank, for instance, whereas cognitive authority is more important when locating scholarly resources. Second, these studies demonstrate that once the purpose of the quality assessment has been narrowed down, people do draw on common criteria to characterize resources. These studies suggest that our goal of identifying common dimensions for assessing the quality of resources for formal classroom use is feasible.

We also examined prior work in machine learning, and the attendant area of annotation studies. Specifically, for a machine learning approach to be feasible, a corpus of annotated resources is needed to train the models. Creating a training corpus requires human beings to annotate each resource as exhibiting, or not, the desired characteristic. The general rule of thumb is that the higher the percentage of agreement between human annotators, the more successful the machine learning algorithm is likely to be. Here, prior research suggests that only approaches that break down a complex judgment like "quality" into clearly defined, multiple dimensions is likely to be successful.

For instance, Devaul and colleagues looked at the inter-annotator agreement on a judgment task that is very much akin to assessing resource quality: deciding if a resource supports a particular educational standard [10]. For this type of broad stroke judgment, inter-annotator agreement was very low, averaging only 32%. Reitsma and colleagues took a slightly different approach to the same standards alignment problem and used a theoretical model to break this challenging human judgment task down into a number of more focused dimensions. They were able to achieve inter-annotator reliability ratings between 61% and 95% for their different alignment dimensions [24]. While we are focused on a somewhat different judgment task (suitability for classroom use versus alignment to a standard), we are taking a similar approach to Reitsma, et. al. However, instead of basing the finer-grained dimensions of human judgments on a theoretical model, we take an empirical approach to identify the most salient dimensions. As our annotation data shows, this empirical approach enabled us to develop a detailed annotation protocol which yielded very high inter-annotator reliability.

A final area of related research that is particularly salient to our approach is prior work in developing similar machine learning models. Some of these efforts seek to examine the correlation between low level features amenable to machine recognition with higher-order human decisions. For example, Ivory and colleagues showed that low-level design issues, such as the amount and positioning of text, or the overall portion of a page devoted to graphics, correlated highly with expert judgments of overall site quality [20]. Other work has explored the feasibility of training models to support quality assessments. Custard and Sumner trained machine learning models to judge overall quality using low-level features like website domain names, the number of links on a page, how recently a page was updated, and whether or not videos or sound clips were present [8]. Their models were able to identify whether a resource was from a high, medium or low quality collection with 76.67% accuracy. Zhu has developed algorithms for improving internet search results that draw on quality metrics such as the time since last update of a web page, presence of broken links, and the amount of textual content [30]. These efforts suggest that our goal of training machine learning algorithms to characterize different dimensions of quality should be tractable. Our research differs from these prior projects in our focus on training models for each individual dimension of quality instead of simply focusing on making a "thumbs-up or down" assessment.

3. CHARACTERIZING QUALITY

We conducted a two part study to identify the characteristics most commonly used by secondary science educators and science librarians when identifying resources for use in the classroom. The first part of the study gathered data from three prior projects where participants were concerned with assessing resource quality for use in educational settings. We reanalyzed these data looking for common criteria or dimensions that participants brought to bear on this complex judgment task. In addition to identifying common high-level dimensions, we also sought to identify the lower-level indicators or features of resources that people draw on to inform their judgments. The first part of this study produced an initial cut at the dimensions and indicators that might comprise a rich, multi-dimensional characterization of resource quality. In the second part, we conducted a mixed-method study of experts with significant experience in selecting online educational resources for classroom use. The purpose of this part was to verify and refine our initial cut of quality dimensions and indicators, and to gather more detailed information on how well (or not) different dimensions and indicators correlated with overall quality judgments.

3.1 Analysis of Prior Studies

To get a clearer picture of how educators characterize quality, an analysis was performed using the raw data collected by three prior projects:

Educator Reviews for DWEL The Digital Water Education Library [12] requires review of their resources by their educational community. We gathered all the educator comments provided in the full reviews for resources targeted at grades 9-12, for a total of 364 reviews generated by 21 reviewers for 182 unique URLs.

Educator Reviews for Climate Change Collection The Climate Change Collection [7] was developed using an interdisciplinary review board for selecting appropriate high quality resources. We obtained all the narrative comments concerning digital resource quality from 55 individual reviews provided by 4 individuals for the 28 grade 9-12 digital resources in the collection.

Educator Focus Groups In 2002, Sumner and colleagues hosted a series of focus groups where science educators discussed the quality of digital library resources [27]. We acquired the transcribed verbal data generated by 38 educators as they reviewed 18 resources.

Name	%
Good general set-up	3.6
Appropriate pedagogical guidance	4.1
Appropriate inclusion of graphics	13.4
Readability of text	10.7
Inclusion of hands-on activities	4.4
Robust pedagogical support	4.1
Age appropriateness	5.3
Appropriateness of activities	2.6
Connections to real-world applications	3.0
Reflects source authority	7.7
Focuses on key content	13.1
Provides access to relevant data	6.1

Table 2: The 12 dimensions of quality accounting for the most comments. The "%" column shows what percent of comments each dimension accounted for.

The qualitative verbal data collected from these three studies was then coded by two raters, first to filter out comments that were not relevant to quality, and then to derive the most important dimensions of quality indicated by the data. The latter was performed in an iterative process where comments were grouped by similarity into categories, and then the categories were iteratively adjusted until they best covered the data. Priority was given to categories that were identified in all three sources of data, and categories were adjusted until 100% interrater agreement was reached.

The results of this analysis were a set of 25 dimensions that described the approaches to quality taken by educators across the studies, and a list of comments in which each dimension was identified. Table 1 gives some examples of such educator comments. Based on the frequency with which each dimension was observed in comments, we selected the top 12 dimensions, as shown in Table 2. These dimensions cover a good range of quality concerns, from technical (e.g. readability and graphics) to content (e.g. presence of activities and key ideas) to contextual (e.g. source identification and real-world applications). The dimensions most often commented on were "Appropriate inclusion of graphics", "Readability of text" and "Focuses on key content".

Thanks to the bottom-up iterative approach taken to identify these dimensions, the study also resulted in a list of low-level *features* or *indicators* for each dimension that identified some concrete and observable factors encompassing the more conceptual dimensions of quality. For example, the "Appropriate pedagogical guidance" dimension had indicators like "Has instructions" and "Identifies learning goals", while the "Age appropriateness" dimension had indicators like "Identifies age range" and "Content is appropriate for age range". Overall, the 12 dimensions above accounted for 78% of all the comments about resource quality in these three studies.

3.2 Expert Study

To verify and refine our dimensions of quality, we performed a mixed-method study of digital library curation experts. We recruited eight digital library experts who also had significant experience as science educators or instructional designers to assess the broader applicability of our initial cut at quality dimensions and indicators. Experts

Sample Dimensions	DWEL Reviews	CCC Reviews	Educator Focus Groups
Appropriate inclusion of graphics	Highly visual site... Provides several interactive visuals that prompt students to think	The movies/graphics are well done, and seem to enhance understanding of the text	I would have liked to see graphics
Readability of text	Does not take advantage of the internet, is difficult to read, and has no illustrations.	Could be formatted differently to be easier to read. The text is dry and technical.	Hard to read with busy background
Focuses on key content	It utilizes various weather sites to introduce concepts to students. Scaffolding of concepts is used throughout.	Looks like a really good activity for seeing... if changes in climate are part of global warming or natural variability.	...lessons looked like they would help convey important information in an engaging way

Table 1: Sample qualitative data drawn from raw data sources for three dimensions.

were presented with a series of digital resources and asked to perform a variety of quality judgments, including the assignment of numeric values to individual quality assessments.

First, each expert thought aloud [17] while evaluating the quality of six digital learning resources. The resources were taken from the Digital Library for Earth System Education (DLESE) [11], a repository of digital educational resources about Earth science. Resources were selected to include both ones that had been peer-reviewed and identified as being high quality, and ones that had been rejected from DLESE for being too low quality. While the experts examined their six resources, their comments about the positive and negative aspects of the resource were recorded. Then they were asked to give the resource an overall rating, from -3 or very low quality, to +3 or very high quality. Finally, they were asked to make an accept/reject decision, that is, they were asked to decide whether the resource was high enough quality to be included in a digital library collection.

Using each expert's ratings of the six resources, three resources were selected for a more detailed review. These resources corresponded to the expert's highest rated resource, the expert's lowest rated resource and the resource closest to the expert's mean rating. The expert then evaluated these resources using the 12 dimensions of quality identified in the analysis of prior studies; again their positive and negative comments were recorded, as well as their ratings of how well each dimension was addressed (from -3 to +3). The products of this study were several hours of verbal data including comments about positive and negative aspects of resource quality, as well as the numerical assessments for overall quality judgments, quality dimension judgments, and inclusion or exclusion from a digital repository.

To assess whether or not the dimensions of quality identified from the previous studies of educators were also used by digital library collections experts, we performed several analyses of these data. We looked at how the dimension level quality ratings compared to both the overall quality ratings and to the accept/reject decisions. Table 3 shows these comparisons. The "Overall" column indicates the correlation between the expert's dimension ratings and their overall resource quality ratings. The "Accept" column indicates the correlation between the dimension ratings and the decision to accept or reject the resource from a digital library. Ten of the twelve dimensions were significantly correlated with overall quality judgments, and 8 of the 12 dimensions were significantly correlated with accept/reject decisions. There were some differences between dimensions most useful for overall quality judgments and those most useful for accept/reject decisions. However, some dimensions were near the top of both lists; for example, "Good general set-up" and "Appropriate inclusion of graphics". These high correlations demonstrate that our identified dimensions of quality are sufficient to capture the subjective processes that digital library experts use when assessing resource quality for classroom use.

Dimension	Overall	Accept
Provides access to relevant data	0.67**	0.46*
Good general set-up	0.65**	0.73**
Appropriate inclusion of graphics	0.61**	0.53*
Robust pedagogical support	0.59**	0.48*
Appropriate pedagogical guidance	0.57**	0.52*
Reflects source authority	0.56**	0.54*
Readability of text	0.54**	0.40
Appropriateness of activities	0.49*	0.53*
Focuses on key content	0.42*	0.32
Age appropriateness	0.41*	0.26
Inclusion of hands-on activities	0.36	0.43*
Connections to real-world applications	0.36	0.27

Table 3: The 12 targeted dimensions of quality and their relationship to digital library experts' assessments of overall quality and accept/reject decisions (*$p < .05$, **$p < .01$). The dimensions are ordered by their correlations with overall quality.

3.3 Indicators of Quality

The expert study confirmed that quality could be decomposed into meaningful dimensions that are more concrete than the abstract concept of "quality". However, even our list of the 12 most important dimensions includes fairly abstract dimensions like "Good general set-up". To make a computational approach to quality feasible, it was necessary to push the decomposition of quality further, identifying low-level *indicators* of quality that were concrete, easily recognizable, and known to be useful for making quality judgments.

Fortunately, candidate indicators for each dimension of quality were already identified in the analysis of previous studies. Thus, the main remaining goal was to identify which such indicators were most important for characterizing quality. To answer this question, we analyzed the verbal data collected from the digital library experts. All of the spoken data recorded during the assessments of overall quality were hand-coded to identify where an expert mentioned a quality indicator as being either present or absent in the

Indicator		Example expert comments of presence (+) or absence (–)
Has sponsor	(+)	I. . . look at the URL just to kind of get an idea of where I'm at, you know, is it like a NOAA site
	(–)	It looks like there's links, but I still don't know who they are.
Has prestigious sponsor	(+)	That is a NOAA site. . . I generally think what NOAA offers up is good quality.
	(–)	If it said NOAA or something like that I would say, "Okay. This is USGS, this is NOAA"
Has instructions	(+)	Well I'm looking and it's talking about how to use the software and I like that
	(–)	Okay, what am I supposed to read here? What am I supposed to start with?
Identifies learning goals	(+)	Again very high quality in terms of what it's addressing and what the objectives are.
	(–)	. . . it would help me if I had some information about where this might fit in the curriculum.
Identifies age range	(+)	Well that's kind of interesting, and it gives grade level which is very nice
	(–)	None of these are really showing me grade level which I'm kind of disappointed
Organized for learning goals	(+)	So, you see how the objectives match up with the worksheets and the procedure.
	(–)	I'm a little unclear as to what's going on. . . and how it connects to the rest of the content.
Content is appropriate for age range	(+)	I think that's a good middle school activity, looking at climate change.
	(–)	I really don't like this already. . . I think middle school kids are probably beyond this.

Table 4: Comments on the presence (+) or absence (-) of quality indicators.

Indicator	Correlation
Has prestigious sponsor	0.905
Content is appropriate for age range	0.889
Has sponsor	0.858
Identifies learning goals	0.842
Has instructions	0.755
Identifies age range	0.728
Organized for learning goals	0.612

Table 5: The seven most predictive indicators and their correlations to accept/reject decisions.

resource. Table 4 shows examples indicators and comments about their presence or absence. When coding was complete, counts for the indicators identified by each of the digital library experts were then tabulated.

Using this data, we examined which indicators were most predictive of the decision to accept or reject a resource. We extracted the indicators where both the presence was highly correlated with acceptance and the absence was highly correlated with rejection. Table 5 shows the top seven such indicators. One of the most reliable indicators of resource quality was the presence of a prestigious sponsor, such as NOAA, NASA or USGS. Other important cues included tailoring the resource content to a specific age range, and giving guidance on using the resource through instructions and identified learning goals.

These indicators provide a concrete definition of quality which corresponds strongly to the expert processes. They provide a set of characteristics that identify the conceptual pieces of a resource that are likely to be considered when judging the quality of a resource. In addition, they provide a means of characterizing quality in terms of low-level features that should be more amenable to computational approaches.

4. COMPUTATIONAL APPROACH

One of the main goals of our research is to computationally assess the presence or absence of the quality indicators in a given resource with an accuracy that approaches human performance. Given the current limitations of computational linguistics, even finding relatively simple indicators, which are trivial to detect for humans, can be exceedingly difficult for automated systems, which still largely lack the ability to *understand* text. However, low-level linguistic tasks such as determining the part-of-speech of words and building syntactic parse trees of sentences have been approached quite successfully; and even on tasks of a deeply semantic nature encouraging results have been achieved (for example the problem of sentiment analysis: given a review of a product, determine if the reviewer generally thought well of the product or not). Similar to such prior efforts, we employ supervised machine learning algorithms to learn a statistical model of the available data. Using such a model, judgments can be made about previously unseen resources.

Supervised machine learning algorithms construct models by statistically analyzing a *training corpus* for which the *correct* judgment is known. To create a model for identifying quality indicators, the training corpus needs to include digital library resources that exhibit an indicator and resources that do not. In order to build this corpus, we ran an annotation project where each library resource was examined for the presence or absence of each of the seven indicators.

Our test bed is the DLESE Community Collection (DCC). The DCC is a collection of interdisciplinary resources within DLESE with a general focus on "bringing the Earth system into the classroom". Criteria for inclusion focus on pedagogical value and accessibility in addition to scientific correctness. Another defining characteristic of the collection within DLESE is that it includes resources that were submitted by individual DLESE users [13]. All submitted resources are currently manually reviewed by committee for compliance with the standards of DCC content. After the decision has been made to include a resource in the collection, it is then annotated with various metadata describing the new catalog entry. In the following sections we describe the protocol and results of the annotation project and the machine learning setup we used to create computational models of the quality indicators, as well as present preliminary results.

4.1 The Annotation Project

We selected 1000 Earth system educational resources directed at high school students; 950 were selected randomly from DCC, and 50 were selected from those resources rejected by DLESE. This uneven distribution reflects the submissions received, which are guided by a very explicit collection scope statement; relatively few resources are submitted that are not good candidates for inclusion. When a reviewer

Table 6: Quality indicator presence in resources and inter annotator agreement

Quality indicator	present in	agreement
Has instructions	39%	85.2%
Has sponsor	97%	99.5%
Has prestigious sponsor	34%	63.6%
Identifies age range	20%	87.3%
Not inappropriate for age	99%	100.0%
Identifies learning goals	28%	83.1%
Organized for goals	76%	80.6%

decides to reject a resource, they write a short free-form note explaining their reasons. Common reasons for rejection, besides quality-related problems, are: the resource is outside the scope of DLESE; the type of the resource is one not cataloged by DLESE; or the resource suggested is already in the catalog. Based on the reviewer's notes we only selected resources that were rejected for what appeared to be quality-related reasons.

Two people with previous experience cataloging DLESE resources were asked to judge the presence or absence of the seven quality indicators on each resource. In order to achieve reliable results we carefully formulated instructions for annotation, outlining our definitions for each indicator using concrete terms and examples taken from our expert study. After a short test-run and in cooperation with DLESE experts we made some minor revisions to these annotation guidelines.

Each annotator was then asked to independently look at 600 of the 1000 resources; they were presented with the home page of the resource and allowed to navigate freely. Every resource was annotated at least once, and 200 were double-annotated to allow us to measure agreement between annotators. If agreement on an indicator is low, it either indicates that the annotation guidelines for that indicator were too inexact, thus letting each annotator come up with their own interpretation, or that annotating that indicator is inherently difficult for people, and their judgment is subjective. It is commonly assumed in *natural language processing (NLP)* that the higher the agreement between annotators, the more likely a machine learning system will be able to approach human performance. Table 6 shows inter-annotator agreement for each indicator, as well as the percentage of resources where the indicator was marked as present. Agreement was above 80% for 6 of the 7 indicators, suggesting that our guidelines were clear and our characterization of quality was not too subjective.

In addition to the annotators' quality indicator judgments, we also recorded the URLs of all web pages that they visited during their review and that they considered to be part of the resource. DLESE only stores the URL of first entry into a resource; but many resources consist of multiple linked pages. Automatically identifying the extent of a resource is a complex problem in itself [16, 14], and one that we're not addressing in this project. Here, we directed the algorithms to only examine the pages considered by annotators to be part of the resource.

After determining the inter-annotator agreement we asked both annotators to cross-check the resources and indicators where they disagreed, and discuss and resolve their disagreement. The resulting set of 200 double-annotated and cross-checked resources can be assumed to be of a higher quality of annotation. These resources were set aside to be used only for the final evaluation of the quality indicator models; following good machine learning practice they are not used during development.

Table 7: Quality indicator predictiveness and leave-one-out analysis

	accuracy
all indicators	71%
w/o *Has instructions*	−5%
w/o *Has sponsor*	−2%
w/o *Has prestigious sponsor*	−16%
w/o *Identifies age range*	−7%
w/o *Not inappropriate for age*	−2%
w/o *Identifies goals*	−4%
w/o *Organized for goals*	−4%

The corpus contains 950 resources that were randomly selected from the DCC, and 50 resources that were not allowed into DLESE for quality reasons; from the perspective of a DCC curator, the 50 rejected resources are of lower quality than the rest. To evaluate if the annotated quality indicators capture aspects of the resources that are relevant to quality assessments, as the expert study suggests, we tested how useful the indicators are in predicting whether a resource was accepted into the DCC. In addition to looking at the utility of all the indicators in combination, we performed a leave-one-out analysis to look at the contribution of each individual indicator. The results of this analysis can be seen in Table 7. These experiments were run on a reduced training set, selected to have an equal number of high quality and low quality resources.

The seven quality indicators together were able to accurately predict whether a resource was ultimately accepted into DCC with an accuracy of 71%. Phrased in another way: given a resource to be reviewed, if we never look at the resource itself, but we know if it has instructions, if it has a sponsor, if it has a *prestigious* sponsor, if it identifies a target age range and isn't clearly inappropriate for it, and if it identifies learning goals and is structured for them, we can predict if it is good enough for inclusion in the collection in about 71% of cases. This is encouraging, as it shows that the quality indicators truly capture relevant aspects of quality. It also leaves room for improvement; an automatic system for assessing quality for a specific task may want to introduce additional indicators to improve performance.

4.2 The Computational Models

The quality indicators are assessed at the level of an entire resource, e.g. an entire resource is considered to either have instructions or not, to be age-inappropriate or not, etc. Thus every classification decision we make looks at a complete resource – containing multiple web pages and possibly rich media and linked PDF files – as a unit.

Machine learning (ML) systems generally operate on numerical vectors, and can't be run directly on collections of textual documents. In order to classify a resource we must encode it into such a vector. The encoding process should attempt to catch salient surface cues (generally called *features*) present in a resource that may help in determining the presence or absence of an indicator while discarding information that will be too complex for the statistical algorithms of the machine learning system. While the support vector machine algorithm we employ has been shown to be very effective at detecting relevant statistical patterns even when the number of features is extremely large (i.e. many thousands), the way in which those features are presented to the algorithm greatly influences how well it will be able to make use of them. Our efforts to find an effective set of features and an effective encoding are guided by a large corpus of previous work in using machine learning on linguistic and semantic tasks; even so the set of features that allow effective models to be built can only be identified experimentally. Here we describe our first iteration towards identifying such features. This implementation is based on the ClearTK toolkit for statistical natural language processing [23].

4.2.1 Feature Extraction

To build the vectorial representation of a resource that is required by the machine learning system, we extract a number of numerical and *yes/no* features of a document. Some features are taken straight from the text (e.g. individual words that show up somewhere in the document); some make use of non-textual elements; other features include the domain the resource is hosted in, or documents that it is connected with (e.g. the sites it links to). The following is the feature set used in the system we are reporting on here:

Bag-of-words This feature set is a common starting point for many natural language processing applications. It simply indicates to the machine learning system if any given word shows up somewhere in the current resource or not. E.g. "resource contains the word 'a' ", "resource contains the word 'seismic' ", "resource contains the word 'record' ", and so on, for every distinct word that a resource contains.[1]

TF-IDF *term frequency – inverse document frequency*, a refinement of the *bag-of-words* feature that gives words a different weight, based on how often they show up in the current resource vs. all resources. For example, the word "and" will show up many times in all resources, so the feature "resource contains the word 'and' " will be indicated to the machine learning system as not very important. On the other hand, the feature "resource contains the word 'Rayleigh' ", assuming the word "Rayleigh" shows up a number of times in the current resource, but almost never anywhere else, will be marked as particularly important.[1]

Bag-of-bigrams Similar to the *bag-of-words* feature, this looks not at single words, but at pairs of words, e.g. "resource contains the word 'long' followed by 'period' ", "resource contains the word 'period' followed by 'Rayleigh' ", "resource contains the word 'Rayleigh' followed by 'waves' ".[1]

Resource URL This feature presents the resource URL to the machine learning system. In addition to the full URL we include the domain and super-domains, e.g. "`http://web.ics.purdue.edu/~braile/edumod/surfwav/surfwav.htm`", "`web.ics.purdue.edu`", "`ics.purdue.edu`", "`purdue.edu`", "`edu`". This helps the machine learning system make useful generalizations about the domain a resource is hosted in.

Table 8: Preliminary Results

Quality indicator	baseline performance	ML performance
Has instructions	61%	78%
Has sponsor	96%	96%
Has prestigious sponsor	70%	81%
Indicates age range	79%	87%
Not inappropriate for age	99%	99%
Identifies learning goals	72%	81%
Organized for goals	75%	83%

URLs linked to We include all the URLs that a resource links to, presented in the same way.

Google PageRank For all URLs we include a feature that indicates the Google PageRank of the respective site. This indicates the relative importance of that site on the internet, measured by how many other sites link to it. For example, a site like `http://www.nasa.gov/` has a high PageRank value, while, e.g. a largely unknown and small university web site will have a low value.

Alexa TrafficRank Alexa[2] is a company offering traffic statistics on web sites based on analyzing user behavior. For all URLs we include their reported *TrafficRank* in our feature set, which indicates the amount of user traffic a web site receives relative to other sites.

4.2.2 The Machine Learning System

The feature extraction process results in one numeric vector per digital library resource. During development the machine learning system analyses these vectors, generated from the training corpus, to learn a statistical model for each of the seven indicators. During evaluation the machine learning system decides if the indicators are present or absent in a resource by applying those models to the vector generated from that resource. We use the SVMlight package [21], which uses the support vector machine approach to machine learning. This approach has been effective in a wide range of natural language processing applications, using features similar to the ones used here. The training parameters are chosen using cross validation: we repeatedly build a model from one part of our training data and evaluate it on the rest, each time refining the parameters of the support vector machine algorithm. The results reported below were achieved using a linear kernel SVM.

4.2.3 Preliminary Results

In order to evaluate the effectiveness of our machine learning setup, we trained and evaluated models on the training data using cross-validation. We then compared the results to a simple baseline: ignoring the resource, and always assuming the most common case. For example, the *has instructions* indicator is present in 39% of resources. If we always assumed that a resource has no instructions, we'd be correct in 61% of cases. An effective machine learning model will show significant improvement over this trivial baseline. Table 8 shows the results of this evaluation.

Good improvements over the baseline were achieved on the *has instructions* and *has prestigious sponsor* indicators, and moderate improvements on the *indicates age range* and

[1] These examples are from `http://web.ics.purdue.edu/~braile/edumod/surfwav/surfwav.htm`

[2] `http://www.alexa.com/`

organized for goals indicators. Using the current feature set we were unable to improve performance over the already high baseline on *has sponsor* and *not inappropriate for age*. Our current features do not appear to be sufficient to determine if a resource identifies its learning goals. These results are encouraging in that even using very basic features we are able to classify some of the indicators fairly well.

5. ERROR ANALYSIS

In order to better understand the current weaknesses and strengths of our models we conducted a study to analyze the errors our system makes. For the purpose of this study we ignored the indicators *has sponsor* and *not inappropriate for age*, because our annotated data provided insufficient variation to conclusively train and evaluate models. We randomly split our annotated training data into a training set (650 instances) and a test set (150 instances), trained quality indicator models on the training set and ran them on the test set. By comparing the models' results with the manual annotation on the test set we identified resources where two or more of the remaining five quality indicator models produced an incorrect result, giving us a total of 39 resources with between 2 and 4 errors each.

A DLESE curation expert who had not taken part in the original annotation project was asked to analyze each of the errors on each of these 39 resources. For each error the expert completed an online questionnaire consisting of both enumerated choice and open-ended questions. This questionnaire asked the expert

- which is correct: the human annotation or the automatic model's result?
- is the indicator clearly present or absent, or is it ambiguous?
- are there cues in the text which clearly signal the indicator and should have been found; or are the cues implied, but not explicitly stated; or are the cues present in graphic elements or other parts not examined by the system (e.g. flash, images, etc.)?

A statistical and qualitative analysis of the results is in progress, and final results are not available yet. However, early results confirm that in 55% of the cases where the model did not detect the presence of an indicator, cues were clearly present in the text and should have been found. As an example, the *Heat Transfer and El Niño* [3] resource clearly lists "curriculum standards", but the algorithms did not recognize these when assessing the *has learning goals* indicator.

In approximately 29% of the cases where the model incorrectly detected the presence of an indicator the expert noted that there was text in the resource which could be mistakenly understood to signal the indicator. For example, for the resource *Fossils in the Field* [4] the machine learning model said that the resource had the *has instructions* indicator. The expert noted that there were parts of the text that superficially may have been mistaken as instructions, stating that "the resource is a discussion of pedagogy" and "the words used provide ideas for instructing students [. . .] but there are no instructions about using *this* resource as a professional reference."

Approximately 15% of the time when the human annotator found the *identifies age range* indicator to be present while the model did not detect it, the cues were buried in graphic elements that are not examined by the algorithms. For example, the *EarthStorm – Relative Humidity & Dew Point* [5] resource lists the grade ranges it supports in clickable buttons and not in the text of the resource.

This kind of information should help us target our future efforts more effectively, and to better understand how quality indicators are encoded within a resource.

6. DISCUSSION & FUTURE WORK

Here we reflect on the efficacy of our methodology, the discriminatory value of the current set of indicators, and future enhancements to the computational models.

6.1 Empirical Methodology

Our work here demonstrates the importance of considering human processes in developing computational models. The high inter-annotator agreement and the encouraging initial performance of our computational models is in part a validation of our empirical methodology for selecting quality indicators. Our expert study took very general dimensions of quality suggested by prior research, and refined these dimensions with digital library experts to produce both qualitative and quantitative data. These data verified that the identified dimensions of quality were in fact used frequently, and perhaps more importantly, these data also allowed us to see how expert attention to particular indicators of quality compared with their decisions to accept or reject a resource. This allowed us to select a set of concrete indicators of quality that were highly correlated with the kind of judgments human experts make. As the computational results showed, by using these empirically derived indicators we were able to make automated methods for characterizing quality more tractable.

6.2 Quality Indicators

We've demonstrated the basic feasibility of our approach for five of the seven indicators. For these five indicators the corpus proved adequate for training models to recognize the indicator with some degree of accuracy. As discussed in the error analysis section, there is still significant room for improvement in these five models.

Vexingly, we were not able to demonstrate progress on two of the seven indicators: *has sponsor* and *not inappropriate for age*. These two indicators show a very uneven distribution in the training corpus, as they are present in almost all of the resources. This is a problem for the statistical processes of the machine learning system, as the algorithms rely on having many examples of both sides (indicator present and absent) to find reliable ways to distinguish between the two cases. In order to make progress on these indicators an extended data set and further annotation will be necessary.

In the research presented here we focused on the seven most predictive indicators from our expert study. We realize that there are additional aspects of a resource's quality that are not covered by these indicators. Potential addi-

[3] http://projects.edtech.sandi.net/roosevelt/elnino/heatandelnino.html

[4] http://www.ucmp.berkeley.edu/fosrec/GrifAll.html

[5] http://earthstorm.mesonet.org/materials/les_rel_humid.php

tional dimensions, such as *inclusion of hands-on activities*, would require further annotations. Other potential dimensions, such as *readability of text*, could be addressed using other methods that do not require further annotation. For instance, Coh-Metrix [15] assesses text cohesion and readability, and there is evidence that its results correlate well with human judgment.

Readers of prior drafts of this document asked why we are modeling both *has sponsor* and *has prestigious sponsor*, when one appears to be a subset of the other. This understandable question arises in part from our poor choice of indicator names; a more appropriate name for *has sponsor* would be *has publisher*. Science educators prefer resources with clearly denoted publishers to ensure that these resources are citable by students, which is a scholarly habit that they are trying to inculcate in secondary science learners. *Has prestigious sponsor*, on the other hand, is more closely tied with assessing the cognitive authority of the source. In short, these two indicators model different judgment tasks and characterize different aspects of resource quality.

6.3 Computational Models

We have presented results that show the feasibility of modeling quality indicators with natural language processing and machine learning techniques. Current results show success on some of the indicators, but performance is still poor on others. Our error analysis shows that the current feature set still misses many cues that, to human readers, are readily available in the text. In order to improve performance across all indicator models we intend to explore a larger set of features that aims to capture the structure of the content and to identify the more important concepts within a resource. In particular we are pursuing two directions:

6.3.1 Surface Structure

Resources that are cataloged by DLESE and other libraries are for the most part in HTML format, potentially linking to PDF files or containing rich media. Currently we use an HTML parser and a simple ad-hoc rule system to extract the text portions of a page, and discard parts that we don't need, such as scripts. The extracted text is noisy: it still contains many things that are not part of the textual content of the web page. In particular it doesn't attempt to distinguish between navigation elements, boilerplate (e.g. page headers or footers, copyright), advertisements, and educational content. A web page offers many visual cues to help the user identify these parts and navigate the text, but in the flat text format we currently use those cues are lost.

We intend to improve on this by building a domain independent system that splits the content into blocks, then classifies those based on textual and HTML cues, to not only identify the non-content parts of a page, but also to split the content into headings and paragraphs. Prior efforts in this area (see for example CLEANEVAL[6]) don't provide the rich structural annotation that we're aiming for and only focus on identifying the textual content.

Having identified those content classes in a resource allows more targeted features. For example, instructions that help the user approach a resource effectively are likely to be found early on in the resource and may be structurally separated from other parts, e.g. consisting of a separate paragraph. With the added information a model for the *has instructions* indicator may be less likely to be distracted by other sections of the resource that use similar terminology.

6.3.2 Semantic Features

The content features used by our system, such as bag-of-words, rely solely on counting words that show up in the training set. This leads to problems when a resource uses slightly different terminology than previously seen resources to describe, for example, learning goals. The automatic system ignores the new words, because they haven't been used when talking about learning goals before; a human reader, on the other hand, could use rich understanding of the words' meaning to recognize that the new words are talking about the same thing. The *Heat Transfer and El Niño* resource mentioned in the error analysis provides a concrete example: the resource referred to learning goals as "curriculum standards", as opposed to other, more common phrases, such as "education standards" or "learning objectives".

On another note, using the current feature set the algorithms see each resource as essentially a large collection of disjointed words, making it hard to distinguish between occasional usages of words like "instruction", and document sections that discuss instructions in a focused way. Lexical methods were used successfully in [9] to identify overarching key concepts within a set of resources. In our future work, we aim to capture more fine-grained, discourse level concepts by using a richer semantic feature set. Having identified key concepts in a paragraph, taking into account the words' semantics rather than just their surface form, the machine learning algorithms should be able to focus on the actual content of the resource versus picking up individual words out of context.

7. CONCLUSIONS

We have presented a principled approach to defining the quality of web resources and training models to perform automatic quality characterizations. Through the analysis of prior work and our own expert study, we identified key *indicators* of quality that are both used by experts in quality assessment and easily recognized by non-experts. We constructed a training corpus of 1000 digital resources annotated with these quality indicators, and trained machine learning models which were able to identify important indicators, like the presence of a prestigious sponsor or age range specifications, with accuracies over 80%. These models can underpin tools ranging from quality assessment engines that can help digital library curators manage large collections to end-user tools that can help students learn to better evaluate the quality of resources they see online.

8. ACKNOWLEDGEMENTS

This material is based in part upon work supported by the National Science Foundation under Grant Numbers CISE/IIS-0741326 and CISE/IIS-0534581. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We also wish to thank our colleagues at Digital Learning Sciences / UCAR for their help with the expert study, the test-bed creation, and the error analysis.

9. REFERENCES

[1] B. T. Adler and L. de Alfaro. A content-driven reputation system for the wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, pages 261–270, Banff, Alberta, Canada, 2007. ACM.

[2] S. ann Knight and J. Burn. Developing a framework for assessing information quality on the world wide web. *Informing Science Journal*, 8:159–172, 2005.

[3] R. G. Baraniuk. Challenges and opportunities for the open education movement: A connexions case study. In *Opening up education: the collective advancement of education through open technology, open content, and open knowledge*, chapter 15. MIT Press, 2008.

[4] J. E. Blumenstock. Size matters: Word count as a measure of quality on wikipedia. In *Proceedings of the 17th International World Wide Web Conference*, pages 1095–1096, New York, NY, USA, 2008. ACM.

[5] T. Carey and G. L. Hanley. Extending the impact of open educational resources through alignment with pedagogical content knowledge and institutional strategy: Lessons learned from the merlot community experience. In *Opening up education: the collective advancement of education through open technology, open content, and open knowledge*, chapter 12. MIT Press, 2008.

[6] CLEANEVAL home page. `http://cleaneval.sigwac.org.uk/`, Oct. 2008.

[7] Climate change collection. `http://serc.carleton.edu/climatechange/`, Oct. 2008.

[8] M. Custard and T. Sumner. Using machine learning to support quality judgments. *D-Lib Magazine*, 11(10), Oct. 2005.

[9] S. de la Chica. *Generating Conceptual Knowledge Representations to Support Students Writing Scientific Explanations*. PhD thesis, University of Colorado, 2008.

[10] H. Devaul, A. Diekema, and J. Ostwald. Computer-assisted assignment of educational standards using natural language processing. Unpublished technical report, Digital Learning Sciences, Boulder, CO, 2007.

[11] Digital library for earth system education. `http://www.dlese.org/`, Oct. 2008.

[12] Digital water education library. `http://www.csmate.colostate.edu/DWEL/`, Jan. 2004.

[13] DLESE Community Collection (DCC) scope statement. `http://www.dlese.org/Metadata/collections/scopes/dcc-scope.php`, Oct. 2008.

[14] P. Dmitriev. As we may perceive: Finding the boundaries of compound documents on the web. In *Proceedings of the 17th International World Wide Web Conference*, 2008.

[15] D. F. Dufty, D. Mcnamara, M. Louwerse, Z. Cai, and A. C. Graesser. Automatic evaluation of aspects of document quality. In *Proceedings of the 22nd annual international conference on Documentation*, 2004.

[16] N. Eiron. Untangling compound documents on the web. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, pages 85–94, 2003.

[17] K. A. Ericsson and H. A. Simon. *Protocol Analysis: Verbal Reports as Data*. The MIT Press, revised edition, Apr. 1993.

[18] B. J. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, and M. Treinen. What makes web sites credible?: a report on a large quantitative study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 61–68, Seattle, Washington, United States, 2001. ACM.

[19] L. Graham and P. T. Metaxas. ”of course it's true; i saw it on the internet!”: critical thinking in the internet era. *Commun. ACM*, 46(5):70–75, 2003.

[20] M. Y. Ivory, R. R. Sinha, and M. A. Hearst. Empirically validated web page design metrics. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 53–60, Seattle, Washington, United States, 2001. ACM.

[21] T. Joachims. Making large-scale support vector machine learning practical. In *Advances in kernel methods: support vector learning*, pages 169–184. MIT Press, 1999.

[22] M. J. Kargar, A. R. Ramli, H. Ibrahim, F. Azimzadeh, and S. B. B. M. Noor. Assessing quality of information on the web towards a comprehensive framework. *Iranian Journal of Engineering Sciences*, 1, 2007.

[23] P. V. Ogren, P. G. Wetzler, and S. Bethard. ClearTK: A UIMA toolkit for statistical natural language processing. In *UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC)*, 2008.

[24] R. Reitsma, B. Marshall, M. Dalton, and M. Cyr. Exploring educational standard alignment: in search of 'relevance'. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 57–65, Pittsburgh PA, PA, USA, 2008. ACM.

[25] S. Y. Rieh. Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 53:145—161, 2002.

[26] B. Stvilia and M. B. Twidale. Assessing information quality of a community-based encyclopedia. In *Proceedings of the International Conference on Information Quality*, pages 442–454, 2005.

[27] T. Sumner, M. Khoo, M. Recker, and M. Marlino. Understanding educator perceptions of ”quality” in digital libraries. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 269–279, Houston, Texas, 2003. IEEE Computer Society.

[28] Wikipedia:featured article criteria. `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria`, Oct. 2008.

[29] H. Zeng, M. Alhossaini, L. Ding, R. Fikes, and D. Mcguinness. Computing trust from revision history. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust*, Oct. 2006.

[30] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 288–295, New York, NY, USA, 2000. ACM.