Characterizing and Predicting the Multi-faceted Nature of Quality in Educational Web Resources

PHILIPP WETZLER, University of Colorado Boulder

STEVEN BETHARD, University of Colorado Boulder

HEATHER LEARY, University of Colorado Boulder

KIRSTEN BUTCHER, University of Utah

SOHEIL DANESH BAHREINI, University of Colorado Boulder

JIN ZHAO, National University of Singapore

JAMES H. MARTIN, University of Colorado Boulder

TAMARA SUMNER, University of Colorado Boulder

Efficient learning from web resources can depend on accurately assessing the quality of each resource. We present a methodology for developing computational models of quality that can assist users in assessing web resources. The methodology consists of four steps: 1) a meta-analysis of previous studies to decompose quality into high-level dimensions and low-level indicators, 2) an expert study to identify the key low-level indicators of quality in the target domain, 3) human annotation to provide a collection of example resources where the presence or absence of quality indicators has been tagged, and 4) training of a machine learning model to predict quality indicators based on content and link features of web resources. We find that quality is a multi-faceted construct, with different aspects that may be important to different users at different times. We show that machine learning models can predict this multi-faceted nature of quality, both in the context of aiding curators as they evaluate resources submitted to digital libraries, and in the context of aiding teachers as they develop online educational resources. Finally, we demonstrate how computational models of quality can be provided as a service, and embedded into applications such as web search.

Categories and Subject Descriptors: H.3.7 [Information Storage and Retrieval]: Digital Libraries; I.2.7 [Artificial Intelligence]: Natural Language Processing

General Terms: Algorithms

Additional Key Words and Phrases: Open Education Resources, Meta-Analysis, Expert Annotation, Machine Learning, Interactive Digital Library Interface

1. INTRODUCTION

We increasingly rely on the World Wide Web to find answers to our questions - Americans alone spend 138 million hours each year searching the web\(^1\) and choosing resources to use. When a user issues a query and is presented with a set of search results containing answers, they need to choose which webpage to use, which resources on a webpage to choose from, what criteria to use when selecting resources, how to combine and/or adapt multiple resources to create a coherent learning experience, etc. It may be cognitively taxing for a non-expert to find what they need in a low quality web page that fails to provide the necessary level of guidance or content structure [18]. In contrast, a high quality web page can make the answer easy to find and comprehend and can therefore help the user quickly and efficiently achieve their learning goal.

However, accurately assessing the quality of a web page and how useful it is to a user’s learning goals can be challenging, especially for a non-expert or beginner. This problem is particularly relevant in the domain of open educational resources (OER), where materials for teaching and learning are made freely available via spaces such as TeacherTube or OpenCourseWare. In OpenCourseWare for example, the processes by which materials are reviewed for quality can differ from one institution to another, and the review of educational materials is often done with a specific context or usage in mind. In such situations, the same educational materials can legitimately be identified as high quality by one reviewer and low quality by another.

One key to addressing this problem is recognizing that the quality of a web resource includes many factors that when put together create a better whole. Quality is a multi-faceted concept, and different facets of quality may be important to different users at different times. For example, a curator of a digital library that is viewing resources submitted to the library might care about the educational reputation of the publishers of the resources, while a teacher that is looking for peer-produced lesson plans on the web might care more that the resources have clear instructions and appropriate learning goals.

The current article aims to better characterize the multi-faceted nature of quality, and to produce computational models of quality that can provide a service to users in the many places that they must assess quality. The key research questions are:

- How can we determine which aspects of quality are most important to users in a particular domain?
- How can we create computational models to identify the presence or absence of such aspects of quality in new web resources?

We find that even a complex concept such as quality can be decomposed into smaller components that are more manageable both for humans and for machines. In particular, we make the following contributions:

- We demonstrate that quality is comprised of several dimensions, and that different aspects of quality may acquire greater focus depending on the needs of the audience. We show how to characterize this kind of multi-faceted quality, both at a coarse level and at a fine-grained level, via a meta-analysis of research on quality and a study of expert processes when assessing quality in web resources.
- We establish a methodology for creating computational models of quality in a particular domain by combining expert knowledge, linguistic annotation theory and machine learning. We show that this methodology is generalizable and produces accurate models for predicting quality in new web resources both in the domain of digital library curation and in the domain of design support for teacher generated content.
- We demonstrate how computational models of quality can be provided as a service, and embedded into applications such as web search. Educators interviewed while exploring this service confirmed the utility of such tools.

2. USE CASES

Making quality choices is a cognitively demanding task. An individual must find and evaluate existing resources and then select those that meet their needs. Individuals often need cognitive support to make strategic choices with online content. Metacognitive and content-based support can be provided in a form that will scaffold the ongoing process.

We envision our computational models of quality being deployed as a service that can support a variety of use cases. Given the multi-faceted nature of quality, different users will often want to
focus on different aspects of quality. The service will therefore take the aspects of quality and the
digital resources that are of interest to the user, and produce annotations of the presence or absence
of the specified aspects of quality in each resource. The following sections describe applications of
such computational models of quality in three specific use cases.

2.1 Curation for Digital Libraries

Educational digital libraries, such as the Digital Library for Earth System Education (DLESE),
would benefit from the use of computational models of quality integrated into their review and
curation processes. These libraries aim to develop high quality collections for interactive teaching
and learning experiences. Materials submitted to DLESE come from various individuals and
institutions, as different materials (e.g. text, videos, simulations), and target different age ranges.
Vetting these materials for quality is a critical and challenging issue for a digital library where
human peer-review of materials is time consuming, expensive, and requires training.

In our experiments, we develop a computational model of quality for DLESE, producing the
underlying technology necessary for developing such assistive tools.

2.2 Support Tool for Peer Production

Within peer-produced environments, such as the Instructional Architect (IA), a simple web-based
authoring tool for teachers, peer-review of the works created is rarely if ever conducted. A
computational model for quality could be used to provide immediate feedback to the creator of the
material, offering potential user interactions that improve the created content. For example, in the
IA system, once a user has finished creating their instructional material, they save it to their
account. The model could automatically run when the material is saved and provide user prompts
indicating deficiencies in the material that would affect its quality such as lack of instructions on
how the material is to be approached or absence of specified learning goals.

In our experiments, we show that our methodology for developing computational models of
quality can also produce models for the domain of Instructional Architect projects.

2.3 Search and Recommendation Interface

Searching for high quality educational materials in major search engines, digital libraries, and
repositories is a time consuming task. For example, if a 9th grade science teacher is looking for a
video about the weather to show students, a simple search in DLESE will yield many results
that the teacher must sort through. We hope that by incorporating machine learning models that assess
the quality of an educational resource a better ranking of search results could be done. This would
reduce the difficulty of sifting through the large number of search results.

In our designed user interface, the information seeker is also able to interact with the algorithm
by changing the relative importance of different aspects of quality to better fit his or her information
needs.

3. LITERATURE REVIEW

Several areas of prior work are relevant to the current project: research on the criteria humans use
when evaluating quality, research on how to annotate quality in web pages, and research in how to
train machine learning models of quality. Each of these areas is discussed in detail below.

When asked to assess the quality of a digital resource or web site, people draw on a large suite of
often implicit criteria to guide their judgment. For example, Fogg [10] found that criteria such as
“ease-of-use”, “expertise” and “trustworthiness” were common among participants in an online
survey on “web site credibility”. Rieh [18] identified criteria such as “accuracy”, “currency”,

...
“trustworthiness”, “scholarliness” and “authoritativeness” as important factors when people were asked to judge “cognitive authority” while searching for information on the web. Sumner [20] found criteria such as “scientific accuracy”, “lack of bias” and “good pedagogical design” to be important in focus groups working with educational digital libraries. Ivory [13] showed that low-level design issues, such as “amount of text”, “positioning of text” and “portion of page devoted to graphics” correlated highly with expert judgments of overall site quality. We use a meta-analysis over such studies to find the most important subset of these many quality criteria.

Given a set of quality criteria, a key question is: do different people apply the same criteria in the same way? Devaul [6] found that in the domain of educational standards, where resources should be annotated as supporting or not supporting a particular educational standard, inter-annotator agreement was very low, averaging only 32%. Reitsma [17] had greater success in this domain by breaking the judgment task down into nine more focused dimensions, including “appeal”, “concepts” and “background”. They were able to achieve inter-annotator reliability ratings between 61% and 95% for their different dimensions. We follow this approach in the current work, which is an extension of [2], trying to break down quality into its key dimensions and indicators before asking annotators to find indicators of quality in digital resources.

Web pages annotated with quality judgments can be used to train machine learning models. A common target of such efforts has been Wikipedia, which offers a large collection of peer produced articles, identifies high-quality “featured” articles, and annotates articles that do not meet Wikipedia quality guidelines. Various teams have tried to predict whether or not an article will be featured, using features such as the presence of complex words, the number of one syllable words, readability indexes such as Flesch-Kincaid or Gunning fog, link structure, Wikipedia categories, etc. [3, 19, 21]. However it turns out that identifying featured articles is quite easy - better articles tend to have more contributors and therefore tend to be longer, so simply predicting that articles longer than 2000 words will be featured achieves 96.3% accuracy [3]. Druck [7] looked at a finer-grained quality task - predicting whether individual edits will be retained or removed in the future of the article, where edits that were retained for a long time were considered high quality, and edits that were quickly removed were considered low quality. Even with a variety of features characterizing the words in the edit, structural changes made by the edit, and article and user edit history, this turns out to be a difficult task - a reverted edit is predicted correctly only 30% of the time.

Machine learning models have also been trained to identify quality in the domain of educational content. Custard [5] trained machine learning models to judge overall quality in the context of educational digital libraries using low level features such as 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. Their models were able to identify whether a resource was from a high, medium or low quality collection with 76.67% accuracy. Aleahmad [1] applied machine learning methods to automatically rate the quality of math solution explanations in the context of developing the QCommons project. Using word features and features about the contributor, their models achieved a 0.67 correlation with the human ratings of quality.

A specific dimension of quality that has received a lot of attention in terms of computational models is readability. A variety of simple measures for readability exist based on the complexity of the vocabulary used, e.g. the Flesch Reading Ease Score [9] and the Flesch-Kincaid Grade Level [15]. Extending these efforts, Graesser [12] designed Coh-Metrix, a tool that computes over 200 measures of text readability covering aspects such as cohesion and syntactic structure. These types of measures have been used as features to train supervised machine learning models in tasks such as assessing the grade level of a text [16] or distinguishing between machine and human generated text [14].
It is also worth making the connection to search in information retrieval, where various measures have been used to identify documents that are not just relevant to the query, but also of high quality. In this context, Google's PageRank algorithm [4], which uses link structure to identify more trustworthy pages, is one metric of document quality. Other quality metrics, such as the time since last update of a web page, presence of broken links, and the amount of textual content have also been shown to improve results in web search [22].

The prior work highlights several important issues: First, there are many potential criteria that can be used to measure quality, and different criteria may be important depending on the specific use case. Second, achieving high inter-rater agreement when judging the quality of web pages may be difficult unless quality is decomposed into smaller, more manageable components. Finally, a wide variety of features based on both page content and link structure have potential for constructing computational models of quality.

4. METHODOLOGY

In the current work, we propose a human centered, replicable approach to producing computational models of quality for a domain of interest. This methodology analyzes studies of quality in related domains, determines how experts understand quality in the domain of interest, and then combines human annotation with supervised machine learning to construct the computational models of quality. Figure 1 illustrates the four steps in our methodology:

1. **Meta-analysis: identifying possible dimensions and indicators of quality.** The goal of this stage is to understand what quality means in the domain of interest. Gather previous studies of quality in the domain of interest and related domains. Identify and code quality-related items (e.g. subject comments, notes) from these studies, and form hierarchical groupings. First group items into high level *dimensions of quality* - coarse groupings of thematically similar items. Then group items in each dimension of quality into low level *indicators of quality* - fine grained groupings with concrete and observable characteristics.

2. **Expert study: selecting key indicators of quality.** The goal of this stage is to select from the identified dimensions and indicators of quality the ones that are most critical to the domain. Enlist experts from the domain of interest to evaluate a small set of example resources based on the dimensions and indicators of quality identified in the previous step. Quantify the expert use of each of the dimensions and indicators, and how to measure their association with overall quality judgments. Identify the most important dimensions and indicators of quality.

3. **Annotation: finding examples of key indicators of quality in example resources.** The goal of this stage is to produce a set of training examples for machine learning models. Gather a varied collection of example resources from the domain of interest. Develop annotation guidelines that show non-experts how to identify the key indicators identified by experts in the previous stage. Enlist a team of non-experts to use these guidelines to identify indicators of quality in the example resources.
4. **Machine learning: training models to predict key indicators of quality.** The goal of this stage is to construct the computational model of quality. Using the annotated collection of example resources, train a supervised machine-learning model for each indicator of quality. Features used by these models should characterize the content of the resource and its relation to other resources, e.g. its words and link structure.

The following sections discuss and evaluate this methodology in three domains: quality in library curation, quality in design support for peer production, and quality in search and recommendation.

5. **EXPERIMENT 1: QUALITY IN DIGITAL LIBRARY CURATION**

In this first experiment, we aim to validate our methodology by applying it to our first use case: predicting the characteristics of quality that are most salient for curators of digital libraries. In
Module 1 Lesson Guide

These Lesson Guides include written and computer activities on concepts of remote sensing in general and NASA's Mission to Planet Earth. More information about Mission to Planet Earth can be obtained on the World Wide Web through NASA's Spacelink site (URL: http://spacelink.msfc.nasa.gov/). Spacelink lists educational materials which are available through NASA's network of Teacher Resource Centers and through NASA's Central Operation of Resources for Educators (CORE), at the address below:

NASA CORE
Lorain County JVS
15181 Route 58 South
Oberlin, OHIO 44074
Tel: (216) 774-1051 (ext. 293/294)
Fax: (216) 774-2144

Section A - What is Mission to Planet Earth?

This is a written activity asking the students to consider what about the earth they would want to study.

Section B - Viewing the Earth from Space

Figure 2: A resource from the DLESE Community Collection (DLESE-000-000-006-612).

In particular, we intend to answer the questions: can a meta-analysis and expert study identify appropriate indicators of quality, and can annotation and machine learning train useful computational models of quality?

5.1 Digital Library for Earth System Education (DLESE) collection

As a basis for this experiment, we use resources from the Digital Library for Earth System Education (DLESE), a digital library whose curators could benefit from assistive tools that analyze the quality of resources submitted to the library. DLESE collects teaching and learning resources about the Earth as a system, and is curated by members of the Earth system education community, including science libraries and the University Corporation for Atmospheric Research (UCAR), who make sure that DLESE resources are relevant and technically stable. DLESE resources may be a single webpage or an entire suite of webpages that should be used together. The front page of a sample resource is shown in Figure 2. DLESE also contains the DLESE Community Collection (DCC), a sub-collection that curators have identified as being resources of the highest quality. Thus, DLESE was a natural place to test our methodology for constructing models of quality - the curators already had a notion of quality, and if we could produce computational models of this, they would help library curators in efficiently processing the resources submitted to the library.

5.2 Meta-analysis: identifying possible dimensions and indicators of quality

Dimensions and indicators of quality were identified from the raw data of three prior projects:
Educator Reviews for DWEL. The Digital Water Education Library 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 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 [20]. We acquired the transcribed verbal data generated by 38 educators as they reviewed 18 resources. 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% inter-rater agreement was reached.

The results of this analysis were a set of 25 high-level dimensions of quality, a set of low-level indicators for each dimension of quality, and a list of comments in which each dimension was identified. Figure 3 gives two example dimensions, the indicators that were nested under each of these dimensions, and an example comment that was associated with each dimension.

Table 1 shows the top 12 dimensions of quality by percent of comments. Overall, these 12 dimensions accounted for 78% of all the comments about resource quality in the three studies.

The output of the meta-analysis was thus a set of 12 high-level dimensions of quality and a set of low-level indicators of quality for each dimension.

---

2 http://serc.carleton.edu/climatechange/
5.3 Expert study: selecting key indicators of quality

To identify which of the low-level indicators of quality were most crucial in the context of DLESE, we performed a mixed-method study of digital library curation experts. We recruited eight experts in digital library curation who also had significant experience as science educators or instructional designers. These experts were presented with a series of digital resources and asked to perform a variety of quality judgments.

First, each expert thought aloud [8] while evaluating the quality of six DLESE resources. While the experts examined their six resources, their comments about the positive and negative aspects of the resource were recorded. Next the experts were asked to give the resource an overall rating, from -3, or very low quality, to +3, or very high quality. Then they were asked to make an accept/reject decision, that is, whether the resource was high enough quality to be included in a digital library collection. Finally, three resources were selected for a more detailed review, and the experts were asked to judge these resources along the 12 dimensions of quality (from -3 to +3) while their positive and negative comments were recorded.

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. The dimension-level judgments correlated well with both the overall quality judgments, and the accept/reject decisions, as shown in Table 1.

We then used the verbal data to identify the key low-level indicators of quality. 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 resource. When coding was complete, counts for the indicators of quality identified by each of the digital library experts were tabulated. We then collected the indicators where both the presence in comments was highly correlated with acceptance and the absence from comments was highly correlated with rejection. Table 2 shows the top seven such indicators.

The result of the expert study was thus a list of seven low-level quality indicators that provide a concrete definition of quality which corresponds strongly to the expert processes. They provide a set of characteristics that identify the conceptual aspects 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.
Annotation guidelines were created for the indicators found in the expert study. This included:

- **Has Instructions**: “A resource that has instructions explains how the content of the resource should be approached by the user. For example, a resource may indicate the sequence in which the user should visit its pages, describe the steps in a classroom activity, or explain how to install and use a piece of software. Simply identifying the contents or purpose of a resource is not considered equivalent to having instructions – there must be some guide that helps understand how to navigate and use the resource.”

- **Has Sponsor**: “A resource that has a sponsor explicitly attributes its content to a person or institution. For example, a page may state that it is organized or maintained by a professor, a
public school, a research lab (e.g. NOAA or NASA), or a politician/celebrity (e.g. Al Gore). Wikipedia-like sites where the content is managed by the community are not considered to have sponsors. Being part of a particular domain is also not sufficient for having a sponsor – the content must be explicitly attributed to someone.”

- **Has Prestigious Sponsor:** “A resource has a prestigious sponsor if the manager or organizer of the page is respected in the field of study relevant to the resource. For earth systems resources, prestigious sponsors could include entities such as NASA, USGS or NOAA, but also respected universities (when the resource is maintained by university itself – not just a student’s webpage).”

- **Identifies Learning Goals:** “A resource that identifies learning goals articulates the knowledge and skills a student is expected to acquire over the course of using the resource. Specific state or national standards may be given, or the resource may more informally identify its best use in an educational setting.”

- **Resource is Organized Appropriately for Learning Goals:** “A resource is organized appropriately for its learning goals if the site is clearly organized so that each goal has a corresponding description or activity. This often means that each learning goal is addressed under a separate heading, tab or page. A page with a single learning goal may be appropriately organized if the headings, etc. give useful sub-structure to the learning activities. If no learning goals were identified by the resource, this indicator should be given the value Not Applicable (N/A).”

- **Identifies Age Range:** “A resource identifies its age range if the text states the expected age or grade level of its intended users, or if the resource is divided into sections targeted to users with different levels of knowledge.”

- **Content Not Obviously Inappropriate for Age Range**: “A resource’s content is not obviously inappropriate for its age range if someone with little expertise in education would judge that the reading or activities were neither much too hard nor much too easy for the given grade level. For example, playing with clay is generally inappropriate for high school audiences, and very technical terminology is inappropriate for elementary school students. In both these cases, the resource should be annotated as No. If no age range was specified by the resource, this indicator should be given the value Not Applicable (N/A).”

---

3 After these studies were completed this indicator was updated to “Content is appropriate for age range” to remove the double negative.
We enlisted as annotators two people with previous experience cataloging DLESE resources but not with explicit judgments of quality. We designed a browser-plugin annotation interface, shown in Figure 4, that allowed the annotators to view a resource in the main panel, while annotating “Yes”, “No” or “N/A” to each indicator in a side panel. Dependent quality indicators were set to “N/A” if the indicators they depended on were marked as “No” (e.g. if Has sponsor was marked as “No”, then Has prestigious sponsor was marked as “N/A”). Annotators were allowed to navigate thorough as many pages of the resource as they felt were necessary to make judgments about the quality indicators, and the pages they identified as being part of the resource were recorded.

The annotators were asked to identify quality indicators for 1000 resources from DLESE, 950 from the DLESE Community Collection and 50 from a set of resources that were submitted to DLESE but rejected by the curators. We selected this subset of DLESE to ensure that there were many examples of good quality indicators, but also negative examples where few or no quality indicators were present. Both annotators annotated a random sample of 200 of the resources so that inter-annotator agreement could be calculated. Table 3 shows this inter-annotator agreement for each indicator\(^5\) in terms of percent agreement, Pearson product-moment correlation coefficient.

\(^4\) After annotator agreement was measured, to produce a final set of official judgments for the collection, the annotators were asked to resolve their differences on these 200 double-annotated resources.

\(^5\) For indicators dependant on another indicator (e.g. Has prestigious sponsor is not applicable if Has sponsor is “No”), agreement is reported only for the cases where the other indicator was present.
Cohen's $\kappa$ and Krippendorff's $\alpha$. The first two metrics do not adjust for the probability of annotators agreeing by chance, while the latter two do. Krippendorff's $\alpha$ reflects expected reliability of the annotation procedure for arbitrary annotators, while Cohen's $\kappa$ reflects only the specific annotators who performed the annotation.

Overall, Table 3 shows that the indicators of quality determined by the meta-analysis and expert study were reliably identified by our annotators. Agreement was good ($\alpha > 0.6$) for Has instructions, Has sponsor, Identifies age range, Identifies learning goals and Organized for goals. For Not inappropriate for age, agreement was perfect, but only because all double-annotated resources were judged as appropriate by both annotators. Agreement was moderate ($\alpha > 0.4$) for Has prestigious sponsor, which requires somewhat more knowledge and understanding of the Earth systems domain.

### 5.5 Machine learning: training models to predict key indicators of quality in DLESE

Having annotated example resources from the collection of interest (DLESE), it was now possible to train machine learning models to predict the quality indicators. We split the annotated DLESE resources into two sets, the 800 single-annotated resources, to be used as training data for the machine learning models, and the 200 double annotated resources, to be used as test data to evaluate the performance of the models\(^6\).

Then for each indicator of quality, we trained a binary classifier on the training data whose task was to predict, given a new resource, whether the indicator was present in that resource or not. We trained Support Vector Machine (SVM)\(^7\) classifiers as they have been found to be effective in a variety natural language processing applications, and scale well to large data and complex problems.

<table>
<thead>
<tr>
<th>Quality indicator</th>
<th>Agreement</th>
<th>Pearson</th>
<th>Cohen's $\kappa$</th>
<th>Krippendorff's $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has instructions</td>
<td>85.2%</td>
<td>0.769</td>
<td>0.675</td>
<td>0.765</td>
</tr>
<tr>
<td>Has sponsor</td>
<td>99.5%</td>
<td>0.721</td>
<td>0.828</td>
<td>0.670</td>
</tr>
<tr>
<td>Has prestigious sponsor</td>
<td>63.6%</td>
<td>0.555</td>
<td>0.316</td>
<td>0.485</td>
</tr>
<tr>
<td>Identifies age range</td>
<td>87.3%</td>
<td>0.739</td>
<td>0.643</td>
<td>0.738</td>
</tr>
<tr>
<td>Not inappropriate for age</td>
<td>100.0%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Identifies learning goals</td>
<td>83.1%</td>
<td>0.696</td>
<td>0.552</td>
<td>0.692</td>
</tr>
<tr>
<td>Organized for goals</td>
<td>80.6%</td>
<td>0.574</td>
<td>0.333</td>
<td>0.691</td>
</tr>
</tbody>
</table>

Table 3: Inter-annotator agreement for each quality indicator on the 200 double-annotated resources.

\(^6\) We reserve the best data for the test set to ensure the reliability of our evaluation.

\(^7\) We used the SVM-light implementation: http://svmlight.joachims.org/
To train these classifiers, we converted each resource, \( R \), into a vector of features, \( V_R \). Each feature of \( V_R \) is a variable that characterizes some piece or measurement of the resource that should be useful for predicting quality. We identified several different feature sets that characterize both the content and the link structure of an educational resource:

- **Bag-of-words**: These features are a common starting point for many natural language processing applications, and frequently the most powerful. They simply indicate the presence or absence of every known word in this resource. Words that occur multiple times are only counted once. For each word that appears in the collection, a binary feature will be added to \( V_R \) that is true if the word was present in the resource and false otherwise.

- **TF-IDF bag-of-words**: These features are similar to the bag of words features, except that instead of indicating the presence or absence of a word with true or false, they try to quantify the importance of that presence by assigning a TF-IDF score. The TF-IDF score is a combination of how frequent the word is in the resource, that is, the term frequency (TF), and how frequent the word is across the entire collection, that is, the document frequency (DF). A word that is useful for characterizing the current resource should be frequent in that resource - have a high TF - and infrequent across the collection - have a low DF, or equivalently a high inverse document frequency (IDF). Formally, given a term, \( t \), a document or web resource, \( d \), and a collection of documents, \( c \), we define \( \text{TF-IDF}(t, d, c) = |\{t' \in d : t' = t\}| * \log(|c| / (1 + |\{d' \in c : t \in d'\}|)). \) So, for example, the word “and” will show up many times in all resources, so the feature “resource contains the word ‘and’” will have a low value. 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 have a particularly high value. Thus, for each word that appears in the collection, a real-valued feature will be added to \( V_R \) that is the TF-IDF score of the word in the resource. Note that this number will be zero for words in the corpus that do not occur in the resource since the term frequency will be zero.

- **Bag-of-bigrams**: These features are similar to the bag of words features, except that instead of indicating the presence of words, they indicate the presence of bigrams, that is pairs of consecutive words. Again, for each bigram that appears in the collection, a binary feature will be added to \( V_R \) that is true if the bigram was present in the resource and false otherwise.

- **Resource URL**: These features present the main URL of the resource to the machine learning system, as a bag of URL segments: the full URL, the domain and the super-domains. Like the Bag-of-words features, a binary feature will be added to \( V_R \) for each URL segment observed in the training data that is true if the URL segment was present in the resource and false otherwise. For example, a resource with the main URL “http://web.ics.purdue.edu/~braile/edumod/surfwav/surfwav.htm” would produce true values for the features “http://web.ics.purdue.edu/~braile/edumod/surfwav/surfwav.htm”, “web.ics.purdue.edu”, “ics.purdue.edu”, “purdue.edu” and “edu”.

URLs linked to: All the URLs that link outward from a resource, i.e. all <a href="..."> elements in the resource’s HTML. Each URL was encoded as a bag of URL segments in the same way as the Resource URL feature.

Google PageRank: For all URLs (Resource URLs and URLs linked to) we included the Google PageRank of the respective site as real-valued feature. This indicates the relative importance of that site on the Internet, measured by how many other sites link to it. For example, a site such as http://www.nasa.gov/ has a high PageRank value, while a largely unknown and small university web site will have a low value.

Alexa TrafficRank: For all URLs (Resource URLs and URLs linked to) we included the Alexa TrafficRank of the respective site as a real-valued feature. Alexa is a company offering traffic statistics on web sites based on analyzing user behavior. TrafficRank indicates the amount of user traffic a web site receives relative to other sites.

We evaluated SVM classifiers learned from the training set\(^8\) against majority class classifiers, which always guess the most common answer for their indicators. For example, the “Has instructions” indicator was present (labeled “Yes”) in 39% of the resources and absent (labeled “No”) in 61% of the resources. Thus, the majority class classifier for “Has instructions” assumes that instructions are absent from all resources, and always produces “No” as its classification.

Table 4 shows the results of this comparison, evaluating on the 200 DLESE resources in the test set. For the indicators Has instructions, Has prestigious sponsor, Indicates age range and Identifies learning goals, the SVM models outperform majority class classifiers by 10% or more (absolute), and reduce the error of the majority class classifiers by about half. For the other indicators, where predicting “Yes” for all items already achieves over 90% accuracy, the SVM classifiers perform the same as the majority class classifiers. These results show that for indicators of quality that are not

<table>
<thead>
<tr>
<th>Quality indicator</th>
<th>Majority class accuracy</th>
<th>SVM accuracy</th>
<th>SVM error reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has instructions</td>
<td>65.2%</td>
<td>88.8%</td>
<td>67.8%</td>
</tr>
<tr>
<td>Has sponsor</td>
<td>98.8%</td>
<td>98.8%</td>
<td>0%</td>
</tr>
<tr>
<td>Has prestigious sponsor</td>
<td>59.6%</td>
<td>80.1%</td>
<td>50.7%</td>
</tr>
<tr>
<td>Indicates age range</td>
<td>83.2%</td>
<td>91.3%</td>
<td>48.2%</td>
</tr>
<tr>
<td>Not inappropriate for age</td>
<td>100.0%</td>
<td>100.0%</td>
<td>0%</td>
</tr>
<tr>
<td>Identifies learning goals</td>
<td>76.4%</td>
<td>90.1%</td>
<td>58.1%</td>
</tr>
<tr>
<td>Organized for goals</td>
<td>92.1%</td>
<td>92.1%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 4: Performance on the 200 DLESE test examples.

\(^8\) For each indicator of quality, we selected the SVM meta-parameters (e.g. kernel type and cost of misclassification) via five-fold cross validation on the training set.
already extremely frequent, SVM models with our set of content and link based features are capable of accurately predicting them.

5.5.1 Feature Importance

To determine which features were most important for identifying which indicator, we performed several analyses. We first grouped features into thematically similar feature sets:

- **URL**: the Resource URL features, with their Google PageRanks and Alexa TrafficRanks
- **Word**: the bag of words features
- **TF-IDF**: the word TF-IDF features
- **Bigram**: the bag of bigrams features
- **Link**: the URLs Linked To features, with their PageRanks and TrafficRanks

We then trained single-feature models, where we try to predict the presence of a quality indicator using only a single feature. If one single-feature model predicts more accurately than another single-feature model, then we conclude that the more accurate feature is more important for detecting that quality indicator. Table 5 shows the results of this analysis\(^9\). The Bigram features were the most important for all quality indicators except for Has prestigious sponsor, where the Link features were the most important. In fact, the Link features were essentially only useful for Has prestigious sponsor - they provided almost no gain over the majority class classifier for all of the other features. This matches our intuition that outgoing links from a resource are good for identifying sponsors, but not very good for identifying more content based quality indicators such as Has instructions or Identifies learning goals.

We next trained models where each feature was removed by itself from the full model. If the performance of the classifier drops dramatically when a feature is removed, then we conclude that the feature was important for identifying that indicator of quality. Table 6 shows the results of this analysis. Most of the losses are relatively small, because there is some redundancy between feature sets (e.g. Words and TF-IDF) so that if one is removed, the model can still compensate by using the other. However, we do observe that removal of the Bigrams features causes a significant drop in performance for the indicator Has instructions, suggesting that word bigrams are important for finding instruction sections. This is consistent with the findings of the single-feature analysis above.

\(^9\) We only include in these analyses the indicators where the SVM classifiers outperformed the majority class models.
Table 6 also shows that performance can actually increase a bit with the removal of a feature set, e.g., up 3.1% (absolute) when the Word features are removed from the Has prestigious sponsor classifier.

In general, the results from the single-feature analysis and feature-removal analysis suggest that Bigram and Link features are important features for quality models, and that there may be some benefit in the future of tailoring feature sets separately for each of the models.

5.5.2 Determining the pages that belong to a resource

Up to this point, we have been evaluating the computational models of quality assuming that we already know which pages belong to which resource. However digital library collections such as DLESE do not in fact contain the resources themselves. Instead, they contain just the primary URLs of resources along with some associated metadata. Importantly, they do not contain a list of all secondary URLs for a multi-page resource. As discussed earlier, we recorded the pages that our annotators identified as being part of the resource during the annotation process, and the results reported above are based on using those manually identified resource pages.

In real world applications of our computational models of quality, all the pages that belong to a resource will typically not be known. Thus it is important to determine how various automatic strategies for identifying the pages of a resource will impact the performance of the models of quality. We therefore compare the following three strategies for crawling and collecting the web pages that belong to a resource:

- **Manual crawl**: Include all the pages of a resource that the annotators visited and considered part of the resource. This will usually only be a subset of the full extent of the resource, since we didn’t ask the annotators to navigate the complete resource; but one would expect that this subset will contain all the relevant information – after all, the annotators didn’t find it necessary to examine any pages outside of this set when making their decisions. This approach requires significant manual effort for each resource.

- **Front page**: Include only the front page of a resource, that is, the page at the URL indicated by the digital library. While this is a very precise approach (we know for certain that this page is part of the resource), the recall is low (many pages that are part of a resource are not included).

- **Depth 1**: Include the front page of a resource, and any pages linked from that front page and within the same domain. This approach will have a higher recall than the Front Page strategy, finding pages beyond just the first one, however it will introduce some errors, for example, if a domain has several separate resources that link to each other from their front pages.
Our goal is not to measure the coverage of these strategies, but instead to measure how the performance of the quality models change when one of these strategies is applied. Thus, each of these strategies was applied one at a time to select the pages in the training set from which the quality SVM models were learned. The resulting models were then evaluated on the test set, and their performance is shown in Table 7. Manually identifying the boundary of a resource results in slightly higher performance than either of the automatic techniques for estimating resource boundaries, but these differences are not statistically significant (using McNemar's test with \( p = 0.05 \)). This is encouraging because it means that to successfully predict indicators of quality, it is not necessary for humans to manually identify the boundaries of the resources.

### Table 7: Performance on the 200 DLESE test examples with three crawling strategies.

<table>
<thead>
<tr>
<th>Quality indicator</th>
<th>Manual crawl</th>
<th>Front page</th>
<th>Depth 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has instructions</td>
<td>88.8%</td>
<td>83.7%</td>
<td>86.1%</td>
</tr>
<tr>
<td>Has prestigious sponsor</td>
<td>80.1%</td>
<td>77.7%</td>
<td>76.5%</td>
</tr>
<tr>
<td>Indicates age range</td>
<td>91.3%</td>
<td>89.8%</td>
<td>88.6%</td>
</tr>
<tr>
<td>Identifies learning goals</td>
<td>90.1%</td>
<td>86.8%</td>
<td>89.2%</td>
</tr>
</tbody>
</table>

Our goal is not to measure the coverage of these strategies, but instead to measure how the performance of the quality models change when one of these strategies is applied. Thus, each of these strategies was applied one at a time to select the pages in the training set from which the quality SVM models were learned. The resulting models were then evaluated on the test set, and their performance is shown in Table 7. Manually identifying the boundary of a resource results in slightly higher performance than either of the automatic techniques for estimating resource boundaries, but these differences are not statistically significant (using McNemar's test with \( p = 0.05 \)). This is encouraging because it means that to successfully predict indicators of quality, it is not necessary for humans to manually identify the boundaries of the resources.

### 6. EXPERIMENT 2: QUALITY IN DESIGN SUPPORT

Having demonstrated that our methodology can produce accurate models of quality on the DLESE collection, we turn to the question of how well this methodology generalizes to other domains by producing accurate models of quality outside the initially studied area. In particular, we are interested in our second use case: how models of quality can support teachers when they are designing educational materials for their classes. We envision such models of quality giving advice to the teachers when their materials need better instructions, learning goals, etc.

Since the same prior work on educator reviews and focus groups would be appropriate for the domain of teacher produced resources, we do not repeat the meta-analysis and expert study steps of the methodology here. Instead, we focus on evaluating the second half of our methodology: given the key indicators of quality, can humans reliably annotate them in teacher generated resources, and can machine learning models accurately predict when indicators of quality are present? This portion of the methodology begins the movement away from expert judgments toward educators and learners and how well the expert judged materials work for learning.

#### 6.1 Instructional Architect (IA) collection

As a basis for this experiment, we use resources from the Instructional Architect (IA), a simple web-based authoring tool for teachers to create open educational resources, which could benefit from models of quality that give immediate feedback on quality issues as resources are being developed. The IA web site provides users with tools to create learning activities (or “projects”) by gathering educational resources from the web, supplementing them with additional text, and publishing the resulting materials to students or the public. A sample IA project is shown in Figure 1. Users of IA would be a natural audience for computational models of quality, as the models could provide real-
time advice for users creating projects, e.g. pointing out when additional instructions or explicit learning goals would improve the quality of the project.

While IA projects are similar to DLESE resources in that they are web sites designed for an educational setting, there are some differences that make this an interesting domain to test the generalization of our methodology. First, IA projects are typically a single webpage, linking out to various other pages, while DLESE resources are often entire sites, providing many pages with detailed information about a particular topic. Second, IA projects are typically produced by individual teachers, not groups or organizations with significant editorial staff as is true for many DLESE resources. Finally, there is no curation team that attempts to identify the highest quality IA projects, where in DLESE such resources are collected in the DCC sub-collection. Thus, if our methodology for producing computational models of quality performs well on the IA collection, then we have managed to bridge some significant domain differences.

6.2 Annotation: finding examples of key indicators of quality in IA resources

Using the key indicators identified by the expert study, and the previous annotation guidelines developed for DLESE, we developed the following annotation guidelines for the IA collection (with comparisons to the corresponding DLESE guidelines noted in parentheses):

- **Has Instructions**: “The IA project includes instructions for the user indicating how to navigate and use the project and linked resources. Example: project indicates a sequence in which the user should visit pages, describes steps in a classroom activity, or explains how to install and use a piece of software.”
  (This is similar to the DLESE indicator but with examples more appropriate to IA.)

- **Links to Prestigious Sources**: “The IA project provides a linked resource(s) to a ‘prestigious’ source. A ‘prestigious’ source includes a site where the manager or organizer is highly respected in the relevant subject area, government organizations, respected digital libraries, university maintained sites, and non-profit organizations. This indicator will require you to click on the resource links. In visiting resources, follow any instructions from the IA project itself (e.g. try to go to the location that students are supposed to go). Once you arrive, do not click on any additional links to make your determination.”
  (The Has sponsor and Has prestigious sponsor indicators are not directly appropriate for Instructional Architect projects, which are authored and published by IA users, generally individual teachers. However, since IA projects generally make extensive use of other resources, and an essential part of their value lies in the web resources that they link to, we asked the annotators to evaluate the “prestigiousness” of the linked sources.)

- **Identifies Learning Goals**: “The IA project identifies learning goals and articulates the knowledge and skills a student is expected to acquire over the course of using the project. Specific state or national standards may be given, or the resource may more informally identify its best use in an educational setting. Remember to use the ‘more information’ link at the bottom of the project as another way to identify learning goals and/or age range.”
  (This is similar to the DLESE indicator, but with specific instructions to follow the ‘more information’ link that is part of the IA template and includes descriptive information that the author entered, including the target age range and a summary of the content.)
- **Organized Appropriately for Learning Goals:** “The IA project is organized appropriately for its learning goals. The goals are clearly organized so that each goal has a corresponding description or activity.”

(This is similar to the DLESE indicator, but modified to account for the fact that all IA projects are contained within a single page.)

- **Identifies Age Range:** “The IA Project identifies its target student age range by stating the expected age or grade level of its intended users in the project itself or in the ‘more information’ link at the bottom of the project; or the project structure, using sections targeted to users with different levels of knowledge, implies a certain age range.”

(This is similar to the DLESE indicator, but again with instructions to follow the ‘more information’ link.)

- **Content Seems Appropriate for Age Range:** “The IA project’s content seems appropriate for its age range; someone with little expertise in education would judge that the reading or activities were neither much too hard nor much too easy for the given grade level. For example, playing with crayons is generally inappropriate for high school audiences, and very technical terminology is inappropriate for elementary school students.”

(This is similar to the DLESE indicator "not inappropriate for age" but rephrased to avoid the double negative.)

We enlisted as annotators three teachers who had previously used Instructional Architect, who had taken part in a professional development workshop within the last two years, who had taught STEM subjects in grades 6-12 for at least three years, and who had used their IA content in their classroom. The annotators used a modified version of the DLESE browser-plugin that used the IA indicators above, and allowed indicator presence to be marked on a five point Likert scale: “strongly disagree”, “disagree”, “neither agree nor disagree”, “agree”, or “strongly agree”. Similar to the DLESE corpus, dependent quality indicators were set to “n/a” if the indicators they depended on were marked as “strongly disagree”. Different from the expert study, this rating scale was used as it is more familiar to the teacher participants than a number scale of -3 to +3.
For annotation, a set of 230 resources from IA was sampled, where each resource contained at least 70 words, referred to at least three external resources, was set to be publicly available, had been viewed at least twenty times, and had been categorized as science, technology, engineering, or mathematics (STEM). All three annotators judged the presence of the indicators of quality for all 230 resources over a period of two months.

Table 8 shows several measures of inter-annotator agreement for each indicator of quality. We calculate Pearson’s product-moment correlation coefficient to evaluate the correlation between annotators, but note that the Pearson coefficient is not entirely appropriate here as it gives credit to annotators who have the same relative ordering even when the absolute values are different (e.g. one annotator that rates 1-2-1-2 against another annotator that rates 2-4-2-4). We also calculate the average (over all pairs of raters) Cohen’s $\kappa$ over the two-class “Yes” vs. “No” decision that was used in DLESE, where we map “strongly agree” and “agree” to “Yes”, “strongly disagree” and “disagree” to “No” and we discard the “neither agree nor disagree” judgments. Finally, we calculate Krippendorff’s $\alpha$ over the full five-point Likert scales.

Table 8 shows that agreement was generally much poorer than that on DLESE, with only the indicator Identifies age range reaching moderate agreement ($\alpha > 0.4$). In inspecting the annotations, we discovered that one of the three annotators used the Likert scale in a very different way: rather than producing a normal distribution over the five point in the scale, they chose never to use the rating “neither agree nor disagree” and seldom used “agree” or “disagree”, preferring “strongly agree” or “strongly disagree” - essentially converting the five point Likert scale into a two point scale. Table 9 shows annotator agreement when this annotator was removed - most indicators now have agreement $\alpha > 0.3$. Nonetheless, we find that agreement is still lower on all indicators in the IA corpus than it was on the DLESE corpus. Two reasons exist for explaining this low agreement among the teachers: 1) teachers struggled to be objective with the projects, viewing the projects through their classroom context and potential use and 2) formal rating of projects was a new task for them.

<table>
<thead>
<tr>
<th>Quality Indicator</th>
<th>Pearson</th>
<th>Two-class Cohen’s $\kappa$</th>
<th>Krippendorff’s $\alpha$</th>
<th>DLESE $\kappa$</th>
<th>DLESE $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has instructions</td>
<td>0.382</td>
<td>0.390</td>
<td>0.297</td>
<td>0.675</td>
<td>0.765</td>
</tr>
<tr>
<td>Has prestigious sponsor</td>
<td>0.363</td>
<td>0.264</td>
<td>0.127</td>
<td>0.316</td>
<td>0.485</td>
</tr>
<tr>
<td>Identifies age range</td>
<td>0.551</td>
<td>0.481</td>
<td>0.417</td>
<td>0.643</td>
<td>0.738</td>
</tr>
<tr>
<td>Appropriate for age</td>
<td>0.312</td>
<td>0.259</td>
<td>0.205</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Identifies learning goals</td>
<td>0.316</td>
<td>0.260</td>
<td>0.271</td>
<td>0.552</td>
<td>0.692</td>
</tr>
<tr>
<td>Organized for goals</td>
<td>0.129</td>
<td>0.185</td>
<td>0.125</td>
<td>0.333</td>
<td>0.691</td>
</tr>
</tbody>
</table>

Table 8: Inter-rater reliability across all three annotators for the IA corpus. The last two columns repeat the reliability numbers on the DLESE corpus for ease of comparison.
Having annotated example resources from the IA collection, it was now possible to train machine learning models of quality. To produce present/absent judgments for each indicator in each resource in the collection, we took the average of the three annotator ratings, then mapped “agree” or “strongly agree” ratings to “present”, “disagree” or “strongly disagree” ratings to “absent” and discarded “neither agree nor disagree” ratings. We then selected two sets from the annotated IA resources, 100 resources to be used as training data for the machine learning models, and 100 resources to be used as test data to evaluate the performance of the models.

For each indicator of quality, we trained an SVM classifier on the 100 training resources using the Word and Link feature sets, and the Front page strategy for identifying resource boundaries. We then evaluated the performance of these classifiers on the 100 testing resources. The first three columns of Table 10 show that our computational models are able to learn to predict the indicators of quality - with around 5% (absolute) better accuracy than the majority class model, and typically reducing the error over the majority class model by more than half. As in DLESE, has prestigious sponsor remained one of the more difficult indicators of quality for the models to identify.

The results in Table 10 allow us to conclude that our methodology for training computational models of quality in our original digital library curator setting (with DLESE) is also able to train computational models for the new setting of supporting teacher generated content (with IA). In both cases, the machine learning models reduce the error to about 50% of that produced by majority class models. This means that both the basic structure of the annotation guidelines and the machine learning features and classifiers appear to be broadly applicable across the different domains.

An interesting follow-up question is whether the models themselves are domain independent. We can investigate this question by asking how well a model trained on the DLESE collection performs

<table>
<thead>
<tr>
<th>Quality Indicator</th>
<th>Pearson</th>
<th>Two-class Cohen’s κ</th>
<th>Krippendorff’s α</th>
<th>DLESE κ</th>
<th>DLESE α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has instructions</td>
<td>0.359</td>
<td>0.373</td>
<td>0.272</td>
<td>0.675</td>
<td>0.765</td>
</tr>
<tr>
<td>Has prestigious sponsor</td>
<td>0.434</td>
<td>0.524</td>
<td>0.366</td>
<td>0.316</td>
<td>0.485</td>
</tr>
<tr>
<td>Identifies age range</td>
<td>0.498</td>
<td>0.365</td>
<td>0.357</td>
<td>0.643</td>
<td>0.738</td>
</tr>
<tr>
<td>Appropriate for age</td>
<td>0.453</td>
<td>0.397</td>
<td>0.351</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identifies learning goals</td>
<td>0.423</td>
<td>0.334</td>
<td>0.372</td>
<td>0.552</td>
<td>0.692</td>
</tr>
<tr>
<td>Organized for goals</td>
<td>0.273</td>
<td>0.526</td>
<td>0.322</td>
<td>0.333</td>
<td>0.691</td>
</tr>
</tbody>
</table>

Table 9: Inter-rater reliability across the two annotators that used the scale normally on the IA corpus. The last two columns repeat the reliability numbers on the DLESE corpus for ease of comparison.

6.3 Machine learning: training models to predict key indicators of quality in IA

The first 30 resources annotated by the annotators were a “training round” and were not used by the machine learning models.
on the IA collection. The last column of Table 10 shows the results of such an evaluation. Only for the indicators Seems appropriate for age and Organized for goals does the model trained on DLESE perform as well as the model trained on IA. And in fact, for most of the other indicators of quality, the model trained on DLESE performs worse than the majority class model (which knows more about the distribution of labels in the IA collection, but makes naive predictions without considering any features of the articles). Thus while our methodology for training computational models of quality generalizes across domains, the specific models trained do not. This is not entirely surprising given that quality is likely to mean different things across different domains.

To try to get an idea about exactly what the differences between the DLESE and IA collections were, we looked at the vocabulary used in the two collections. DLESE contains many specialized scientific terms from the field of Earth science, and while some scientific terms also show up in the IA collection, only 57% of words that occur five or more times in DLESE show up in the IA collection at all. However, if that 57% of words contains a common set of terms used when presenting indicators of quality, then it might be possible to focus the machine learning models on those words, and thereby get models trained on DLESE that still perform reasonably well on IA.

We attempted to limit the models to such core words in two different ways. First, we filtered the words in the DLESE examples down to just the words that were common to both DLESE and IA. Second, we filtered the words in the DLESE examples down to words that are common in general English writing. To identify such words, we used the Brown corpus, which is a large multi-genre collection of written English including news coverage, political and industry reports, and fiction [11]. Filtering by IA produces a very small set of words which might be most appropriate to predicting quality on IA, while filtering by the Brown corpus aims at a more general set of words and allows the model to put some weight on words that don’t occur in IA if they are very predictive for DLESE. After the word filtering, SVM models were trained on DLESE as before.

Table 11 shows the results of evaluating these models on the IA test set. Filtering down to only the words in both DLESE and IA hurt the model, while filtering down to the words in both DLESE and the Brown corpus generally improved the model. However the gap between even the Brown-

<table>
<thead>
<tr>
<th>Quality indicators</th>
<th>Majority accuracy</th>
<th>IA-trained accuracy</th>
<th>IA-trained error reduction</th>
<th>DLESE-trained accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has instructions</td>
<td>92.4%</td>
<td>98.0%</td>
<td>73.6%</td>
<td>89.8%</td>
</tr>
<tr>
<td>Has prestigious sponsor</td>
<td>53.0%</td>
<td>61.0%</td>
<td>17.0%</td>
<td>44.0%</td>
</tr>
<tr>
<td>Indicates age range</td>
<td>91.0%</td>
<td>96.1%</td>
<td>56.7%</td>
<td>12.6%</td>
</tr>
<tr>
<td>Seems appropriate for age</td>
<td>93.0%</td>
<td>98.9%</td>
<td>84.3%</td>
<td>98.9%</td>
</tr>
<tr>
<td>Identifies learning goals</td>
<td>91.3%</td>
<td>97.7%</td>
<td>73.6%</td>
<td>50.0%</td>
</tr>
<tr>
<td>Organized for goals</td>
<td>91.2%</td>
<td>96.8%</td>
<td>63.6%</td>
<td>96.8%</td>
</tr>
</tbody>
</table>

Table 10: Performance on the 100 IA test examples, using either majority class classifiers, the classifiers trained on the IA collection, or the classifiers trained on the DLESE collection.
filtered DLESE model and the model trained on the IA training set was still quite wide. Thus, the DLESE and IA collections are not just using different vocabulary, but they are using their different vocabularies in different ways. In particular, the words and phrases that indicate quality in DLESE are not the same as the words and phrases that indicate quality in IA. This again suggests that while the methodology for creating models of quality generalizes across different collections, different models will be needed to characterize quality in different domains.

7. CASE STUDY: QUALITY IN SEARCH

One way to use machine learning models that can predict the quality of educational resources is to use them in the ranking and presentation of search results. The most straightforward way to incorporate quality prediction into ranking is to rank the results based on their overall quality score. To test this, we trained 7 classifiers, one for each quality indicator. The overall quality score for a resource is produced by averaging over the scores of individual indicators. However we can enable faceted search of educational resources by looking at the scores for individual quality indicators instead of combining them. For example, we can rank resources based on how they score on the ‘Has Instructions’ indicator only. Indeed any possible combination of indicators can be used for ranking.

We expect that this kind of faceted search is especially beneficial to K12 teachers searching for supplemental materials for classroom teaching. Ranking by individual quality indicators can potentially assist teachers searching for resources. It allows teachers to search for resources based on one quality indicator and assists them with at a glance scanning of materials. This could save the teacher’s time with resource acquisition.

In order to allow users to incorporate quality information in their search we modified an existing educational search interface for the DLESE collection. The modified UI, shown in Figure 6, allows the user to switch the ranking scheme from the default method to one based on overall quality score. The user is also able to rank using individual or any subset of the seven quality indicators.

We also used the quality prediction models for extracting snippets of text that could serve as summaries in search result presentations. To achieve this, instead of inputting whole resources into the quality prediction models, we broke the resource up by paragraphs which were used as input into the models. Paragraphs that scored the highest were then chosen to be presented as representative snippets of a search result in the search UI.

To evaluate the use and feasibility of this UI, during a conference open session teacher educators provided verbal (n=15) feedback on the importance and usefulness of the system. At the “Open Science Fair” portion of the conference, one laptop computer displayed the search interface. Participants were asked to perform a simple keyword search of their choosing related to earth science. The search results displayed the quality indicators with the resources (see figure 5). Participants were also asked how they saw this search feature being used in K12 educational settings. Each educator (n=15) who searched and viewed the results was excited about the possibilities the search and results process presented for K12 teachers. When the participants were asked if the quality indicators should be visible to the user or run in the background, the majority of them (n=12) indicated that having the quality indicators visible (marked as present or not present) in search results would provide a quick scan of the materials. In turn, this would save time in evaluating an online resource for classroom use.
8. SUMMARY

In this article, we have demonstrated a successful methodology for characterizing and predicting the multi-faceted nature of quality in web resources: a meta-analysis decomposes quality into a hierarchy of high-level dimensions and low-level indicators, an expert study identifies the key low-level indicators of quality for a particular domain, annotators manually identify the presence or absence of indicators in example web resources from the domain, and machine learning models examine the content and link structure of the annotated resources to learn to predict quality indicators in new web resources.

We showed that the methodology that developed models of quality in the context of digital library curation (as examined by our DLESE use case) could also build models of quality in the context of teachers generating educational materials (as examined by our IA use case). The meta-analysis and expert study produced a characterization of quality that was appropriate across the two use cases, identifying seven key indicators of quality to be annotated in DLESE, six of which were easily applied to IA. Similarly, in the machine learning step of the methodology, the same type of supervised classifiers with the same set of content and link features yielded computational models in both DLESE and IA that could predict indicators of quality with accuracies of 80% or higher.

We have also demonstrated how such computational models of quality can be provided as a service, and integrated into applications such as web search. Educators interviewed while exploring this service confirmed the utility of such tools.

One of the limitations of this study was that, though our meta-analysis identified 25 dimensions of quality and many indicators of quality under each dimension, we had only the resources to annotate example web resources and construct models for the 7 indicators of quality that were identified as being most important by the expert study. Though the same set of features might work for some of the indicators of quality we did not address, there are also many indicators where we would likely need new features - for example, “Good balance between graphics and text” would require features identifying the size of images in the resource, and “Writing style is readable and understandable” would require drawing in features from the prior work on readability.

<table>
<thead>
<tr>
<th>Quality Indicator</th>
<th>DLESE</th>
<th>DLESE IA-filter</th>
<th>DLESE Brown-filter</th>
<th>IA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has instructions</td>
<td>89.8%</td>
<td>82.7%</td>
<td>90.8%</td>
<td>98.0%</td>
</tr>
<tr>
<td>Has prestigious sponsor</td>
<td>44.0%</td>
<td>45.8%</td>
<td>47.5%</td>
<td>61.0%</td>
</tr>
<tr>
<td>Indicates age range</td>
<td>12.6%</td>
<td>10.7%</td>
<td>17.5%</td>
<td>96.1%</td>
</tr>
<tr>
<td>Seems appropriate for age</td>
<td>98.9%</td>
<td>98.9%</td>
<td>98.9%</td>
<td>98.9%</td>
</tr>
<tr>
<td>Identifies learning goals</td>
<td>50.0%</td>
<td>36.0%</td>
<td>51.1%</td>
<td>97.7%</td>
</tr>
<tr>
<td>Organized for goals</td>
<td>96.8%</td>
<td>95.8%</td>
<td>95.8%</td>
<td>96.8%</td>
</tr>
</tbody>
</table>

Table 11: Performance on the 100 IA test examples, for the original DLESE-trained models, two models where word features are filtered by external corpora, and the IA-trained models.
Another limitation was in the ability of humans to reliably agree on how to annotate indicators of quality. Though we found high levels of inter-rater agreement on indicators of quality in the DLESE use case where the annotators were digital library catalogers, we found much lower levels of inter-rater agreement in the IA use case where the annotators were teachers. In the latter case, formally rating resources was a new experience for the teachers, and different teachers applied the same rating scale in different ways. Fortunately, we found that even with the lower levels of agreement between the annotators, by aggregating their judgments we were still able to train useful computational models of quality.

Finally, while the methodology for constructing computational models of quality easily generalizes from one use case to another, the individual models trained via the methodology often do not. Only two of the models trained on DLESE performed as well on the IA test data as the models built from the IA training data. Analysis of these models showed that though humans were able to characterize quality using the same set of indicators for both DLESE and IA, the words and phrases that identified these quality indicators in the text had little vocabulary in common between the two collections, and thus the models trained on DLESE fared poorly when evaluated on IA.
In the future, we intend to extend and replicate this work by identifying key indicators of quality in domains beyond Earth systems (e.g. biology, engineering) and for other educational levels (e.g. higher education). In particular, this would involve replicating the meta-analysis and expert study in these new domains to see if there are significant differences in the facets of quality that are most important in each domain.

We also plan several extensions in the construction of the machine learning models. As we start to tackle new indicators of quality, we will almost certainly need to incorporate new types of features, such as measures of readability, multi-media structure or semantic cohesion. We have also seen evidence in the current study for trying to select the most appropriate subset of features for each indicator of quality before training the machine learning models.

Finally, building on the user interface for searching and the feedback from educators, we plan to explore new ways to integrate the quality models into a variety of web applications. In the context of digital libraries, they can serve as the basis for both better search rankings and for a faceted search that helps users find resources that have the indicators of quality they care most about. In the context of peer-production, the models can serve as immediate feedback to the creator of a web resource, making specific suggestions on how to improve the quality of the content.

REFERENCES


