This is a preprint of an article published in Library & Information Science Research: Stvilia, B., Al-Faraj, A., & Yi, Y. (2009). Issues of cross-contextual information quality evaluation—The case of Arabic, English, and Korean Wikipedias. *Library & Information Science Research*, *31*(4), 232-239.

Issues of Cross-Contextual Information Quality Evaluation—The Case of Arabic, English, and Korean Wikipedias

Besiki Stvilia[1], Abdullah Al-Faraj, and Yong Jeong Yi

College of Communication and Information, Florida State University

101 Louis Shores Building, Tallahassee, FL 32306, USA

{bstvilia, aka06, yjy4617}@fsu.edu

Abstract

An initial exploration into the issue of information quality evaluation across different cultural and community contexts based on data collected from the Arabic, English, and Korean Wikipedias showed that different Wikipedia communities may have different understandings of and models for quality. It also showed the feasibility of using some article edit-based metrics for automated quality measurement across different Wikipedia contexts. A model for measuring context similarity was developed and used to evaluate the relationship between similarities in sociocultural factors and the understanding of information quality by the three Wikipedia communities.

Introduction

There is consensus that quality is contextual and dynamic. The set of criteria that define better or worse quality can vary from one context to another (Strong, Lee, & Wang, 1997; Stvilia, Gasser, Twidale, & Smith, 2007). With growing efforts by government, business, and academia (e.g., Federal Bureau of Investigation, 2009; TeraGrid, 2009) to aggregate different kinds of data from different sources, there is an increasing need to develop a better understanding of the effects of different contextual relationships (cultural, social, economic) on data integration and management in general and on information quality (IQ) management in particular. To measure IQ effectively, it is essential to be able to model the general context of information use and the components of the context, and to define how changes in these components may affect the perceived quality.

Research in IQ measurement aggregation is related to many long-standing and still active research issues in knowledge representation, information integration, and computational linguistics (e.g. Batini, Lenzini, & Navathe, 1986; Halevy et al., 2005). When aggregating IQ measurements from heterogeneous data sources from different contexts, one may need to attend to differences in the concept trees of IQ measurement models, vocabularies, metric functions, and measurement representations with regard to scale, precision, and formatting. One may also need to identify differences in the value structures, reference sources, and requirements for quality between the source and destination contexts.

Culture is one of the main structures of an activity context, along with community practices and social and technological factors. According to activity theory (Kuutti, 1991; Leontiev, 1978), a continuous

[1] Corresponding author

feedback loop exists between an activity and its context. Cultural artifacts and factors (e.g., language, beliefs, norms, conventions) interact with and transform an activity and its components. A conceptual tool such as a quality model also could be influenced and shaped by the sociocultural relationships of its use.

With encyclopedias in more than 150 languages and its communities—along with the logs of editorial and quality-assurance processes—open to the general public, Wikipedia provides an excellent opportunity for studying issues of cross-contextual IQ evaluation, and measurement reuse. This study examined community understandings and models of IQ in Arabic and Korean Wikipedias, and then compared those with each other and with the quality model used in English Wikipedia. The researchers also looked at ways of measuring differences or commonalities at different levels of the information creation context (e.g., culture, community) and their effects on IQ evaluations.

Problem statement

There has been a growing emphasis on the need for information reuse and integration in academia, government, businesses and online communities (e.g. Atkins et al, 2003; Library of Congress, 2008). At the same time, context sensitive information processing, service provision, and quality management remain important research and practical problems of information management and cyberinfrastructure design (Gasser, Sanderson, & Zdonik, 2007). To manage IQ effectively, it needs to be measured first. Each measurement, however, has a cost associated with it. Also, data managers and users may not be always able to assess directly the quality of information they aggregate (e.g. due to the lack of access or subject knowledge), and have to relay instead on quality assessments and rankings supplied by local information providers. For reusing or adapting quality measurements to a particular context, it is essential to be able to measure similarities and differences between different contexts and to understand how they are related to quality measurements. For example, when aggregating IQ measurements, data managers may need to attend not only to differences in conceptual IQ measurement models, vocabulary, metric functions, and measurement representations with regard to scale, precision, and formatting, but may also need to identify differences in value structures, reference sources, requirements and constraints in quality between the source and destination contexts of aggregation. Finally, identifying differences between the source and destination context may help determine what quality-assurance interventions will be required and be effective for a particular quality dimension (e.g. Completeness, Accuracy).

Several general and information type-specific IQ assessment models have been proposed in the IQ literature (e.g., Bruce & Hillman, 2004; Eppler, 2003; Fallis & Frické, 2002; Stvilia, 2007; Stvilia et al., 2007; Wang & Strong, 1996). No research has been done, however, on issues of cross-contextual quality measurement and measurement reusability. This study aimed to address that gap.

Literature review

The IQ of an information object can be evaluated either directly, by examining the object itself, or indirectly, by analyzing records of the object's creation and mediation or use processes, and/or the reputation of its creator(s) (Stvilia et al., 2007).

A number of studies have looked at the quality of Wikipedia from different perspectives. Lih (2004) studied Wikipedia content construction and use processes from the perspective of participatory journalism. Lih proposed two measures of quality: Rigor (i.e., Total Number of Edits) and Diversity (i.e., Total Number of Unique Editors). Emigh and Herring (2005), on the other hand, used distributions of linguistic features in the content of the article to compare the formality of language between two community-based encyclopedias (Wikipedia and Everything2) and the Columbia Encyclopedia. Stvilia, Twidale, Smith, and Gasser (2005) developed a model for measuring the IQ of English Wikipedia articles. In that study, a set of metrics was proposed that could be measured automatically. The set

consisted of 11 measures based on the article edit history metadata (e.g., total number of edits), as well as eight measures representing the article attributes and surface features (e.g., readability scores). The measures were then factor analyzed to extract an underlying model with seven constructs: (1) Authority/Reputation; (2) Completeness; (3) Complexity; (4) Informativeness; (5) Consistency; and (6) Volatility. Wilkinson and Huberman (2007) found a positive correlation between the number of edits and the quality of Wikipedia articles, after controlling for the variables of article age and popularity. Other studies, however, showed that the relationship between the number of edits and the quality of an article might not be monotonic. For controversial articles, a high number of edits could be an indicator of conflicts and edit wars, and could be negatively correlated with the quality of the article (Kittur, Suh, Chi, & Pendleton, 2007). Furthermore, as the size of the article's editorial group increases, the cost of coordination can be detrimental to quality. Kittur and Kraut (2008) found that an increase in the number of editors had a positive effect on quality only when work was concentrated within a small group of editors, and had a negative impact when the work was more evenly distributed among editors.

Human activities are mediated and transformed by artifacts, and artifacts are shaped and evolve through use and interaction with other components of the activity and its outer sociocultural context (Kuutti, 1991; Leontiev, 1978). Bryant, Forte, and Bruckman (2005) interviewed nine editors to identify how the sociotechnical structures, including community norms and conventions, mediated and shaped user activity in Wikipedia, and how the users' motivation and roles evolved over time. Viegas, Wattenberg, Kriss, and van Ham (2007) revisited their earlier study and analyzed how different English structural variables of the English Wikipedia had evolved over time. Similar to another earlier study (Stvilia, Twidale, Smith, & Gasser, 2008), they found that the work coordination artifacts of Wikipedia had been growing and becoming increasingly sophisticated.

The literature provides many definitions of culture (e.g. Geertz, 1973; Kroeber & Kluckhorn, 1952). Most often culture is defined as shared symbolic forms that people and communities use to express meaning and which also shape their behavior and social processes. These symbolic forms include language, stories, rituals, norms, beliefs, and art forms (Swidler, 1986). Also, culture is learnt through interaction (Axelrod, 1997; Star & Ruhleder, 1996). Whereas the above definition of culture places emphasis on social artifacts associated with culture, the definition of culture proposed by Hofstede underlines the social cognition aspect: "the collective programming of the mind that distinguishes one group or category of people from another" (Hofstede & McCrae, 2004, p. 58). Hofstede's name is also associated with a widely used model for measuring cultural differences among nations (Hofstede, 1991; Hofstede & McCrae, 2004). The model includes five constructed dimensions based on the results of large-scale surveys conducted in different countries over a period of time: (1) Power Distance Index (the level of the society's endorsement and acceptance of social inequality); (2) Individualism (the degree to which individuals are integrated in groups—the strength of group ties); (3) Uncertainty Avoidance (the level to which the society protects its members from uncertainties by enacting laws, rules, security measures, and sets of beliefs); (4) Masculinity (the degree of difference in emotional roles among sexes); and (5) Long-Term vs. Short-Term Orientation (whether the society values thrift and perseverance more than it respects tradition and social obligation).

A number of studies have applied Hosfstede's cultural dimensions to Wikipedia. Pfeil, Zaphiris, and Ang (2006) examined possible relationships between Hofstede's cultural dimensions and distributions of edit action types in French, German, Dutch, and Japanese articles on the topic "Game." The researchers found that the probability of editors performing different kinds of actions in the Wikipedias was correlated with differences in scores on the cultural dimensions. However, because the analysis used only four articles on a single topic, whether this finding is generalizable to the Wikipedias is not known. In addition, the differences in the edit type distributions could have been influenced by noncultural factors, such as articles being at different stages in their life cycles.

In addition to cultural factors, variances in the socioeconomic characteristics of Wikipedia communities might lead to differences in editor participation and edit patterns. Rask (2007) compared

participation rates in 12 different language Wikipedias with the values of the corresponding Human Development Indices. The author found that more developed countries exhibited a higher rate of participation in Wikipedias than underdeveloped countries. Like Pfiel et al. (2006), Rask too hypothesized that cultural factors may affect how openly editors from different cultures might express their opinions, and suggested that a correlation could exist between Hofstede's cultural dimensions and the edit types in different versions of Wikipedia.

Furthermore, it has been suggested that cultural differences might affect the quality of knowledge representation artifacts. Hammwöhner (2007) looked at the consistency of cross-language article linking and subject categorization among English, French, German, and Italian Wikipedias. This analysis of the classification trees used suggested that the percentage of shared or "international" editors could be correlated with the level of consistency of subject classification and indexing among of the corresponding articles on those Wikipedias. The study did not test the significance of the correlation, however.

In an earlier study, Stvilia and Gasser (2008) identified the following sources of IQ change: changes in Culture, Community, Activities, Agents, Knowledge, and Technology. Culturally related changes might include changes in language, norms, or conventions. In addition, interactions and relationships among these different levels of IQ change could be 'many to many'. For instance, a community's IQ decision making could be influenced and mediated by multiple cultures, and an activity system might use multiple technologies or tools. Alternatively, the same culture could be shared by many communities, and the same agents and tools could be shared by more than one activity system. Thus, by inference, an increase in the variance in these factors might lead to lower perceived quality at the local level, and sharing cultural, community, or technology artifacts might result in higher perceived quality at the global level (see Table 1).

Table 1. Sources of IQ change[a]

Culture	The culture changes—what was admissible and aligned with the value system of the previous culture might not be admissible or interpreted in the same way in the current culture.
Community	The community makeup as a whole changes—it could become smaller or larger, more aligned or less aligned, more selective or less selective.
Activities or Events	New activities could be introduced that might generate new needs and uses for the information object. Alternatively, some of the existing activities in which the information object was used might become obsolete, making the related information needs obsolete as well. New events might occur that could affect the information object directly (e.g., initiation of a peer-reviewing or quality assessment process) or indirectly through its underlying entity (e.g., a country has elected a new president).
Agents	Changes occur in editorial groups—existing editors could leave or become inactive; new editors could arrive who might not be aligned with the group, might be less qualified, or might not be interested in contributing faithfully (e.g., trolls, spammers).
Knowledge, Technology, or Tools	The current state of knowledge changes—what was considered accurate in the past might not be accurate now. New technologies could be developed that might change the cost structure for activities, including quality assurance activities—activities that were prohibitively expensive in the past might become affordable now. Alternatively, a tool or technology might become less effective or efficient with the changed reality, or might simply malfunction.

[a]Stvilia and Gasser (2008).

Research questions and hypotheses

The study used the general frameworks of IQ assessment and change proposed earlier (Stvilia & Gasser, 2008; Stvilia et al., 2007) and a literature analysis to formulate the following research questions:

- What quality models are used by the Arabic, English and Korean Wikipedias?
- Is there a relationship between the quality models used by the Arabic, English and Korean Wikipedias?
- Are quality evaluations or measurements transferable from one Wikipedia to another?
- Do Wikipedia communities with more similar socio-cultural characteristics have more similar quality models?

In addition, analysis of the related literature suggested that the following hypotheses be tested:

H_1: The Wikipedias have different models of quality.

H_2: The *Number of Edits* is positively related to quality.

H_3: The *Number of Unique Editors* is positively related to quality.

H_4: Wikipedias that have more similar cultural characteristics are likely to have more similar quality models.

Procedures

Wikipedia quality assurance work is highly dynamic and consists of at least six kinds of processes: (1) those that create and maintain Wikipedia articles; (2) those that evaluate the quality of an article and act on it directly by modifying, deleting, or changing its status; (3) those that evaluate the performance of Wikipedia editors and select quality assurance agents (administrators, automatic scripts [bots]); (4) those that identify and block malicious editors; (5) those that coordinate and manage editor activities, and resolve conflicts; and (6) those that build and maintain Wikipedia's work coordination and articulation artifacts. A more detailed description of the Wikipedia work organization can be found elsewhere (Stvilia et al., 2008).

Featured articles (FA) are identified by the community as Wikipedia's best articles. Articles can be nominated as candidates for FA status by individuals or a group. Once nominated, the candidate articles undergo a peer review process and are voted on to determine whether they meet the Wikipedia FA criteria. In particular, reviewers examine the FA candidate article for compliance with Wikipedia's FA criteria and comment if any problem is noted (hereafter referred to as an IQ problem). Reviewers also indicate their support for or opposition to promotion of the candidate article to FA status. At the end of the peer review process, based on the number of votes, an FA director or another administrator in charge of the FA process decides whether to promote the candidate to FA status. Thus, it is important to note that the FA promotion or demotion processes are not completely horizontal. The FA process rules and procedures, and in a few cases, final decisions as well, have been significantly influenced by Wikipedia administrators (Stvilia et al., 2008). Nevertheless, both published (formal) quality evaluation guidelines and, more importantly, an implicit or active model of quality exhibited through FA votes and decisions can shed light on a particular understanding of IQ in a Wikipedia community and the value structure for that IQ.

The study used copies of the June 2008 dumps (copies) of the Arabic, English, and Korean Wikipedia databases. The dumps contained 37,583, 2,632,551 and 74,321 articles respectively. From each Wikipedia the study sampled and analyzed 1,000 random articles, 60 FA candidate articles and votes, 10 user pages, and FAs from each Wikipedia. The study also screen scraped edit history entries for all articles in the sample. Half the FA candidate vote samples contained promoted articles, and the other half included rejected articles. The FA samples consisted of 91 Arabic articles, 1,047 English articles, and 25 Korean articles. These differences in size among the FA samples were caused by differences in the total number of FA articles in the three Wikipedias.

The study was guided by activity theory (Kuutti, 1991; Leontiev, 1978), and by the model of IQ change developed in an earlier study (Stvilia & Gasser, 2008; see Table 1). These theories suggested areas and levels of variability, as well as points of interaction within and across the different Wikipedias. In particular, the researchers looked at how the different levels and components of the context of an information activity differed or coincided across the Wikipedias, and how these differences or commonalities were related to IQ judgments. Hofstede's cultural dimension scores (ITIM International, 1967–2009) for the United States, South Korea and the Arab World were used to assess pair-wise similarity of the Wikipedias at the cultural level. Article edit based measures (e.g. *Number of Edits, Number of Editors, Number of Edits*) can be considered as indirect representations of a Wikipedia community's editorial activities. Hence, the study used article edit-based measures to evaluate similarities between the Wikipedias in terms of community makeup (e.g. registered vs. anonymous editors) and editorial activities.

To answer the research questions and test the hypotheses, a mixed set of methods was used, including descriptive and multivariate statistics, content analysis, and similarity measures from the fields of information retrieval (Salton & McGill, 1982) and information theory (Cover & Thomas, 1991). In particular, to identify differences and similarities among the formal IQ models of the Wikipedias, the FA selection criteria and policies were examined and contrasted. Actual practices—the uses of FA criteria – were also examined. Logs of FA candidate votes were content analyzed to identify what virtue(s) or lack thereof (quality problems) Wikipedia editors referred to when promoting or rejecting a particular FA nomination or candidate.

Furthermore, to identify the structure of IQ assurance activities, the set of actions performed by Wikipedia's quality assurance agents (administrators) was examined. The user and discussion pages of 10 randomly selected administrators were collected from each Wikipedia and were content analyzed for the IQ assurance actions requested by editors.

To quantify the similarities or differences among the three Wikipedias similarity metrics from information retrieval and information theory were used. The frequencies of quality problem types found in each vote were normalized to 1 or 0. The edit based measures, however, evaluated related but different concepts (e.g., *Number of Internal Links* vs. *Number of Edits*) and certain measures also had different scales (e.g., *Number of Adjacent Pages* vs. *Number of Registered Users*). To compensate for these differences, the researchers used the Cosine Angular similarity metric, which measures the "angle" between vectors, rather than the distance between points, in a vector space (e.g., the Euclidean distance measure). Likewise, to measure similarity at the cultural level, the Cosine similarity metric was applied to the Hofstede's cultural dimension scores. In addition, the Jaccard similarity metric was used to evaluate the degree of subject overlap among the article collections.

In the content analysis, each researcher independently coded the data for his or her area of responsibility by using an open-coding procedure. The codings were compared to ensure intercoder reliability, and where there were inconsistencies, the researchers resolved these as a group; these decisions were then were incorporated into subsequent coding exercises. The resultant codes were iteratively clustered to develop final coding schemas (Bailey, 1994). These final schemas were then used to recode the samples. At the end of the analysis of the FA candidate vote logs, the researchers mapped the final coding schemas onto the set of IQ dimensions from the IQ measurement framework (Stvilia et al., 2007).

Finally, to test the research hypotheses, multivariate statistical analysis was applied to the results of the content analysis of the FA candidate vote logs and the values for article edit-based quality metrics (e.g., *Number of Edits*, *Number of Editors*).

This is a preprint of an article published in Library & Information Science Research: Stvilia, B., Al-Faraj, A., & Yi, Y. (2009). Issues of cross-contextual information quality evaluation—The case of Arabic, English, and Korean Wikipedias. *Library & Information Science Research*, *31*(4), 232-239.

Findings

The researchers began by looking at formal (published) policies and guidelines related to quality. The formal FA criteria or "quality virtues" of the Arabic Wikipedia were an exact match with the formal set of criteria from the English Wikipedia (see Figure 1). The Korean Wikipedia, however, had fewer quality dimensions. Interestingly, the Korean quality criteria set did not include the dimensions of *Stability* and *Comprehensiveness*/*Completeness*.

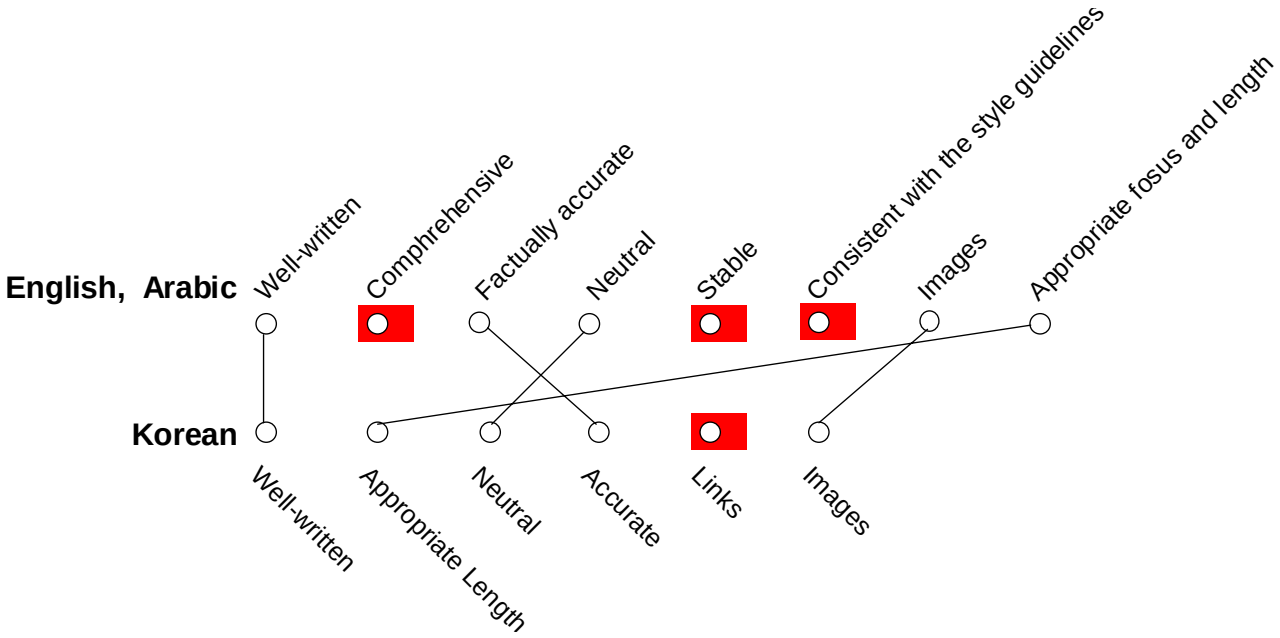

Figure 1. The IQ criteria mapping (the lines are used to indicate a semantic match between the criteria; the red rectangles indicate that the criteria do not have semantic match in the other Wikipedia).

In addition to the differences in the formal models of quality, the Wikipedias exhibited different active IQ models, as determined by the kinds of IQ problem claims made by FA candidate reviewers. The analysis of IQ problem distributions from the FA candidate votes and the relationships between the IQ problem types and the vote outcomes (i.e., promoted or rejected) showed statistically significant differences among the Wikipedias. A Kruskal-Wallis correlation test revealed that almost all the IQ problem types, with the exception of *Neutrality* and *Volatility*, were significantly dependent on the version of Wikipedia (see Table 2).

Table 2. Kruskal-Wallis test of dependence of the occurrence of IQ problem claim types (normalized to 1) on the version of Wikipedia (Arabic, English, or Korean)

	Accuracy	**Cohesiveness**	**Completeness**	**Consistency**	**Image**
Chi-square	21.25	7.33	32.06	51.86	43.19
df	2	2	2	2	2
Asymp. sig.	0.000	0.026	0.000	0.000	0.000

	Prose	Neutrality	**Relevance**	**Verifiability**	Volatility
Chi-square	53.77	3.34	20.04	11.35	1.07
df	2	2	2	2	2
Asymp. sig.	0.000	0.189	0.000	0.003	0.586

In addition, the Wikipedias differed in their value structures for quality. A binary logistic regression (logit) of the quality problem types on the outcomes of the FA candidate votes revealed that the Wikipedias appeared to place importance on different quality virtues or the lack of those virtues. The logit model for the Arabic Wikipedia (model fit likelihood ratio: $\chi^2 = 16.77$; $p < 0.02$) showed the *Accuracy* and *Appropriate Length* problem claims to be significant predictors of FA candidacy rejection. A similar analysis of the Korean Wikipedia (model fit likelihood ratio: $\chi^2 = 15.84$; $p < 0.03$) identified the *Accuracy* and *Prose* (*Well-Written*) quality problems as statistically significant predictors of negative quality judgments. In the English Wikipedia, however, none of the problem types was a statistically significant predictor of FA candidacy rejection.

In the next step, the researchers regressed the quality problem types on the version of Wikipedia (i.e., language version) using multinomial logistic regression. The English Wikipedia was used as a baseline. The regression analysis confirmed (model fit likelihood ratio: $\chi^2 = 176.18$; $p < 0.0001$) that the problem claim type distributions in the Arabic and Korean Wikipedias differed significantly from the type distributions in the English Wikipedia. The analysis revealed that the presence of *Accuracy*, *Completeness*, *Consistency*, *Image*, and *Prose* problem claims increased the odds of the FA candidate being from the English Wikipedia, compared with the Arabic and Korean Wikipedias ($p < 0.05$). Likewise, having *Prose* (*Well-Written*) and *Verifiability* problem claims increased the odds of Wikipedia being Korean, compared with Arabic ($p < 0.05$). Thus, the English and Korean Wikipedia communities had a more detailed or nuanced active model of IQ than the Arabic Wikipedia.

Kruskal-Wallis tests of the FA candidate and random article samples showed significant dependence ($p < 0.001$) of most of the article edit-based measures (*Number of Editors*, *Number of Edits*, *Number of Adjacent Pages*, *Number of Edits by Registered Users*) on the version of Wikipedia.

In addition, several earlier studies of the English Wikipedia suggested that the *Number of Editors*, *Number of Edits*, and *Number of Edits by Registered Users* could be used as indirect measures of the quality of an article (e.g., Lih, 2004; Stvilia et al., 2005; Wilkinson & Huberman, 2007). To test whether these measures had the same relationships with quality in different Wikipedias, the study used logit analysis to regress the values of the measures for the FA candidates on the outcomes of the votes. The analysis revealed that only the logit model of the Korean Wikipedia was statistically significant (model fit likelihood ratio: $\chi^2 = 14.33$; $p < 0.003$). Furthermore, although all three measures maintained the same sign for the relationship across the three Wikipedias, only the *Number of Editors* and the *Number of Edits* were significant predictors of FA promotion, and only in the Korean Wikipedia sample (see Table 3).

Table 3. Relationships between edit-based measures and FA promotion (180 FA candidates). The '+' sign indicates a positive relationship between the measure and FA promotion decision, while the '-' sign means that an increase of the measure decreases the odds of FA promotion.

Metric	Arabic	English	Korean
Number of Editors	–	–	–[a]
Number of Edits	+	+	+[a]
Number of Edits by Registered Users	–	–	–

[a]Indicates that the measure is a statistically significant ($p < 0.05$) predictor of FA promotion in a particular Wikipedia version.

A multinomial logistic regression of the aggregate set of edit-based measures of FA candidates on the three Wikipedias (model fit likelihood ratio: $\chi^2 = 176.18$; $p < 0.0001$) showed that increasing the *Number of Edits* increased the odds of an FA candidate being from English Wikipedia compared with the Arabic or Korean Wikipedia ($p < 0.01$). However, the *Number of Edits by Registered Users* was a positive indicator of the article being from the Arabic and Korean Wikipedias ($p < 0.01$), compared with the English Wikipedia.

A multinomial logistic regression analysis of the random samples of articles from the three Wikipedias showed somewhat different relationships among the measures and the version of Wikipedia. In contrast to the FA candidates, random Arabic and Korean articles had a smaller *Number of Editors* and *Number of Edits by Registered Users* but a larger *Number of Edits* than English articles (Table 4). In addition, the randomly selected articles from the Arabic Wikipedia had a smaller *Number of Edits by Registered Users* (or a larger *Number of Edits by Anonymous Users*) but a larger *Number of Editors* than the Korean articles. Furthermore, the English Wikipedia articles were more sparsely connected to each other through shared editors than were the Arabic and Korean articles—an increase in the *Number of Adjacent Pages* decreased the odds of the article being from the English Wikipedia.

To quantify the similarities (or differences) between the extracted IQ models and the cultural characteristics of the Wikipedia communities, the researchers used similarity metrics from information retrieval. In particular, the Cosine Angular similarity metric was used, and was calculated as follows:

$$\text{Cosine}(x, y) = \frac{\sum_{i=1}^{t} d_{ix} \times d_{iy}}{\sqrt{\sum_{i=1}^{t} d_{ix}^2} \times \sqrt{\sum_{i=1}^{t} d_{iy}^2}},$$

where *x* and *y* indicate the Wikipedia models (quality or culture), and *d* is the individual dimension or criterion. In the active quality models, *d* is a normalized frequency of a particular quality IQ problem claim type, whereas in the cultural models, *d* stands for a Hofstede dimension (ITIM International, 1967–2009).

The analysis showed that the Cosine similarity score between Korean culture and U.S. culture was lower (i.e., the cultures were more dissimilar) than the cosine scores for the other two pairs of Wikipedias, Arabic and English, or Arabic and Korean (see Table 5). Because the Hofstede criteria website did not list a score for the Long-Term Orientation dimension for the Arab World, this dimension was not used in the calculation.

To assess similarity at the community level, the researchers used cosine similarity scores for the following article edit-based measures: *Number of Editors*, *Number of Edits*, *Number of Internal Links*, *Number of Edits by Registered Users*, and *Number of Adjacent Articles*. In particular, 500 articles were randomly selected from each Wikipedia, and values on the measures for each article from one sample were matched with values on the measures for every article in the other sample, resulting in 250,000 pairs of value sets for each Wikipedia pair. Next, multinomial logistic regression was applied to cosine similarity scores for the value set pairs to explore the relationships among the Cosine similarity score distributions and the Wikipedia pairs, as well as to produce similarity-based rankings of the pairs. The

regression analysis (model fit likelihood ratio: $\chi^2 = 74{,}839$; $p < 0.0001$) showed a significant dependence of the Cosine similarity scores on the pairs of Wikipedias. The analysis revealed that the Arabic and Korean Wikipedias were the most similar to one another at the community level. That is, an increase in the value of a Cosine similarity measure increased the odds of the pair being of the Arabic and Korean Wikipedias compared with the other pairs of Wikipedias ($p < 0.001$). In addition, the analysis showed that the Korean and English pairs had higher Cosine similarity scores than the Arabic and English pairs (see Table 5).

Similarly, multinomial logistic regression analysis of the active quality models (model fit likelihood ratio: $\chi^2 = 940.37$; $p < 0.0001$) showed that the Wikipedia pairs differed significantly in their Cosine similarity score distributions ($p < 0.001$). However, for the IQ models, the Korean and English Wikipedias had the most similar IQ problem claim distributions, followed by the Arabic and English pairs and the Arabic and Korean pairs.

Table 4. Multinomial logistical regression test of dependence of the edit-based measures on the Wikipedia version (Arabic, English, or Korean)[a]

Measures	RRR	Std. Err.	*z*	*P* > *z*
Arabic/English				
Number of Editors	0.936	0.013	–4.77	0.000
Number of Edits	1.029	0.008	3.73	0.000
Number of Adjacent Pages	1.054	0.003	18.87	0.000
Number of Internal Links	1.007	0.002	4.00	0.000
Number of Edits by Registered Users	0.918	0.010	–7.79	0.000
Korean/English				
Number of Editors	0.870	0.013	–9.19	0.000
Number of Edits	1.028	0.009	3.18	0.001
Number of Adjacent Pages	1.055	0.003	19.22	0.000
Number of Internal Links	1.009	0.002	5.41	0.000
Number of Edits by Registered Users	0.942	0.011	–5.13	0.000
English/Arabic				
Number of Editors	1.069	0.015	4.77	0.000
Number of Edits	0.972	0.007	–3.73	0.000
Number of Adjacent Pages	0.949	0.003	–18.87	0.000
Number of Internal Links	0.993	0.002	–4	0.000
Number of Edits by Registered Users	1.089	0.012	7.79	0.000
Korean/Arabic				
Number of Editors	0.930	0.013	–5.3	0.000
Number of Edits	0.999	0.006	–0.21	0.833
Number of Adjacent Pages	1.001	0.001	1.74	0.082
Number of Internal Links	1.002	0.001	2.21	0.027
Number of Edits by Registered Users	1.026	0.009	2.85	0.004

[a]Model fit likelihood ratio: $\chi^2 = 1{,}654.39$, $p < 0.0001$; sample size = 3,000 articles.

This is a preprint of an article published in Library & Information Science Research: Stvilia, B., Al-Faraj, A., & Yi, Y. (2009). Issues of cross-contextual information quality evaluation—The case of Arabic, English, and Korean Wikipedias. *Library & Information Science Research*, *31*(4), 232-239.

Table 5. Pair-wise similarity rankings for the Wikipedias[a]

Levels of comparison	1	2	3
Culture	Arabic Korean	Arabic English	Korean English
Community	Arabic Korean	Korean English	Arabic English
Active Quality Model	Korean English	Arabic English	Arabic Korean

[a]Cosine angular similarity measure was used. For each level of comparison, the pairs are arranged in decreasing order of similarity, from 1 to 3.

The content analysis of FA votes showed a desire by the Wikipedia communities to encourage members to write and promote articles on local topics and priorities. A few editors in Arabic Wikipedia expressed dissatisfaction with having articles on non-Arabic topics nominated for FA status:

> "We need encyclopedia articles that interest Arabic readers. . . . I wish you have made this effort to write about a subject that benefits your people."

Indeed, the percentage of FA candidates on topics of local importance in the Arabic Wikipedia was only $9/60 \times 100 = 15\%$, whereas in the Korean Wikipedia, this number was $13/60 \times 100 = 21\%$. Despite this dissatisfaction, the statistical analysis (Kruskal-Wallis test) did not reveal a significant relationship (influence) between topic locality and FA vote outcomes or results in either the Arabic or Korean Wikipedia.

The content analysis also suggested that a considerable exchange of content had taken place among the Wikipedias:

> "Just look at the English version of this article and you will know why [the article cannot be FA]."

> "I just saw that this is a featured article—it would be nice if the major parts could be ported to the English Wikipedia. From there it could find its way into the other Wikipedias as well. :)"

The analysis of editor pages revealed translation as one of the most frequently performed tasks, particularly in Korean Wikipedia. Analysis of the random samples showed that the subject overlap (i.e., having articles on the same topic) with English Wikipedia was higher in Korean Wikipedia than in Arabic Wikipedia (59 vs. 48%). In addition, the Korean Wikipedia articles had a significantly higher *Number of Shared Editors* with English Wikipedia than did the Arabic Wikipedia articles. Differences on the Jaccard similarity scores for the *Number of Shared Editors* with the English Wikipedia was not statistically significant, however (see Table 6). Also, the portion of shared contributors in the total *Number of Editors* was not high in either Wikipedia: 9% in the Arabic Wikipedia, and 13% in the Korean Wikipedia. The Jaccard similarity scores were calculated as follows: $(N \cap E)/(N \cup E)$, where N refers to the editor set in the non-English (Arabic or Korean) Wikipedia sample and E refers to the editor set of the English Wikipedia version.

Table 6. Kruskal-Wallis test of dependence of the *Number of Shared Editors* with the English Wikipedia on the language version for the Korean and Arabic Wikipedias[a]

	Jaccard scores for Number of Shared Editors	Total Number of Editors in non-English versions	Number of Shared Editors with English version
Chi-square	3.32	2.94	8.94
df	1.00	1.00	1.00
Asymp. sig.	0.068	0.086	0.003

[a]The combined set of random samples was 2,000 articles. Only 352 articles (174 Korean and 178 Arabic) shared editors with the English Wikipedia. The 178 Arabic articles had total of 1345 editors and 133 shared editors. The 174 Korean articles had total of 964 editors and 144 shared editors.

Discussion

The understanding of IQ is context dependent. The same information could be evaluated as being of different quality in different contexts. Based on an analysis of the previous literature, a set of research questions and hypotheses was developed to investigate the relationship between context change and IQ evaluation in Wikipedia.

The first research question sought to identify what IQ models were used by the different Wikipedias and how those models were related to each other. In an effort to answer this question, both formal (published) and latent models of quality extracted from FA candidate vote logs were analyzed. Results showed that differences among the extracted (active) models of IQ of the Wikipedias were statistically significant and thus supported the first hypothesis (H_1). The Wikipedia communities differed both in the set of quality virtues by which articles were examined and in the weight placed on those virtues when making FA promotion decisions. The values for article edit process measures also showed a significant dependence on the version of Wikipedia.

The second research question was concerned with identifying measures that could be used to assess IQ directly or indirectly, and whether IQ measurements were transferable from one Wikipedia context to another. The researchers looked at three article edit-based measures: the *Number of Edits*, the *Number of Editors*, and the *Number of Edits by Registered Users*. For the FA candidates, results showed that the measures were consistent in their relationship with FA promotion decisions (see Table 2). The *Number of Edits* was positively related to quality (FA promotion) in all the Wikipedias, thus supporting the second hypothesis (H_2). This result is also in line with some earlier studies on the English Wikipedia, which showed that the *Number of Edits* was positively correlated with IQ (Wilkinson & Huberman, 2007). The third hypothesis (H_3), however, was not supported. The *Number of Editors* and the *Number of Edits by Registered Users* were negatively related to FA promotion. Although earlier studies of the English Wikipedia (e.g., Lih, 2004) suggested that the *Number of Editors* could be positively related to quality, more recent studies by Kittur et al. (2007) and Kittur and Kraut (2008) showed that the relationship between the *Number of Editors* and quality in the English Wikipedia was more nuanced. They found that an increase in the size of an article's editorial group had a positive impact on quality only when a small core of editors performed the majority of editorial work. Future research examining the dynamics between the *Number of Editors* and IQ could help define optimal and critical values for the measure for each Wikipedia. The optimal value would define the level above which the benefit of adding an editor would be lower than an increase in the coordination cost.

The analysis of the random samples showed that the distributions of edit-based measures of the random samples were different from the distributions of FA candidates. The randomly selected Arabic and Korean articles had a higher *Number of Edits* but a smaller *Number of Editors* and *Number of Edits by Registered Users* than the randomly selected English articles (Table 4). If the *Number of Edits* were accepted as a "global" indirect indicator of quality, then the findings could be interpreted as meaning that randomly selected articles from the English Wikipedia were of lower quality than randomly selected articles from the Korean and Arabic Wikipedias. A manual evaluation and comparison of the IQ of randomly selected articles from the Wikipedias could provide more insight into this issue and could help confirm or reject this hypothesis.

The third research question was concerned with the relationship between the degree of similarity in cultural characteristics and the degree of similarity in IQ models. The findings were not that conclusive, as in the case of the first question. Although the analysis showed that the Arabic and Korean Wikipedias as were the most similar at the cultural and community levels, the active IQ models of the Korean and English Wikipedias emerged as the closest to each other (see Table 5). Thus, the fourth hypothesis (H_4) was not supported.

The study used the edit based measures to evaluate similarity of the Wikipedias at the community level. At the time of analysis, the Arabic and Korean Wikipedias had similar sizes (i.e., within the same scale or order of magnitude) of article collections, whereas the number articles in the English Wikipedia was almost two orders of magnitude greater. In addition, the English Wikipedia articles were more sparsely connected to each other through shared editors than were the Arabic and Korean Wikipedia articles. The finding that the Arabic and Korean Wikipedias were more similar at the community level suggests a possible relationship between the size of the Wikipedia database and community, and the structure and information work patterns of the community. Also, further research is needed to investigate whether and how differences in the size of the article databases and the size of communities might influence the IQ models of the communities and their decision making.

Results revealed that the Arabic Wikipedia had the same formal model for quality as the English Wikipedia. In practice, however, the Arabic Wikipedia was less similar to the English Wikipedia than the Korean Wikipedia on the distribution of the quality virtues the communities used to examine FA candidates. The difference could be related to the Korean Wikipedia sharing a higher percentage of editors with the English Wikipedia than the Arabic Wikipedia. A qualitative study focusing on international editors or "boundary crossers" could offer additional insight into the role of those editors and their effect on the convergence of understanding of IQ among different cultures and communities.

Furthermore, in a previous study, Rask (2007) found a correlation between the Human Development Index and Wikipedia participation rates. Possible future research might investigate whether differences in the active quality models and differences in the community level measures (i.e. the edit based measures) could be related to differences in socioeconomic variables of the nations associated with these Wikipedias.

Finally, the study showed that substantial subject overlap and content sharing were taking place among the Wikipedias. At the same time, more than 40% of articles in the Korean Wikipedia and more than 50% of articles in the Arabic Wikipedia were on topics that did not have matching versions in the English (global) Wikipedia. This finding points to the importance of non-English Wikipedias as valuable sources of knowledge and information. It also underscores the need for mechanisms of cross-Wikipedia IQ evaluation, selection, and aggregation.

The study has certain limitations. A convenience sample of only three Wikipedias was examined, out of more than 150 existing Wikipedias. The selection of these particular Wikipedias for the analysis was determined by the language expertise available in the research team. In addition, when exploring the communities' understanding of IQ and value structure for IQ, the researchers relied solely on content analysis of the FA vote logs - one of many IQ assurance processes and decisions in the Wikipedias. Furthermore, each sample was coded by a single coder proficient in a particular Wikipedia language. Use of multiple coders would increase the reliability of the results. To measure the cultural similarity among the Arabic, English, and Korean Wikipedias, the researchers used Hofstede's cultural dimension scores for the Arab world, the United States, and South Korea (ITIM International, 1967–2009). None of these scores represents a complete set of nations that speak a particular language used in Wikipedias. In addition, it is not known how well Hofstede's scores represent the cultures of open, online communities such as Wikipedia, where there are no geographic boundaries and editors from different cultures can contribute to different Wikipedias. It is important to note that the article edit-based measurements used in this study were generated by screen scraping and parsing the edit history entries of the articles. The information extracted from these history entries is only an approximation of actual editor identities. It is not necessary for an individual to be logged on or even registered to make an edit. In addition, the same individual can be registered and make edits under more than one name within the same Wikipedia or different Wikipedias. Agreement between the current research findings and results of prior research points to the validity of this work. However, replicating the study with larger samples and with different Wikipedias would strengthen the results.

Conclusion

This study represents an initial exploration into the issue of IQ evaluation across different cultural and community contexts. These findings provide valuable insight into the understanding of IQ by three Wikipedia communities. Results of the study also contribute to a better understanding of the problems of automatic IQ evaluation, IQ measurement reuse, and measurement of information context similarity. The study findings can benefit different classes of information providers, intermediaries and end-users (e.g. data managers, librarians, journalists) for whom information and metadata quality evaluation, maintenance, and quality based information selection are essential tasks.

Results showed that different Wikipedia communities may have different models for quality. Results also showed the feasibility of using some of the article edit-based metrics in automatic measurement of IQ across different Wikipedia contexts. Future work may include defining information and community type-specific profiles consisting of IQ models, metric definitions (along with baseline and critical values), and adaptors for IQ context and measurement translation and aggregation.

The study also developed a hierarchical model for context similarity measurement. The Hofstede cultural dimensions were used to represent the cultural level of the model. Future work may include extending the model by adding additional socio-economic indices (e.g. Human Development Index), and testing it with different samples of information systems.

References

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., Messina, P., Ostriker, J. P., & Wright, M. H. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure.* Retrieved March 25, 2009, from http://www.nsf.gov/oc/oci/reports/atkins.pdf

Axelrod, R. (1997). The dissemination of culture: a model with local convergence and global polarization. *The Journal of Conflict Resolution, 41*(2), 203-226.

Bailey, K. (1994). *Methods of social research* (4th ed.). New York: The Free Press.

Batini, C., Lenzini, M., & Navathe, S. (1986). A comparative analysis of methodologies for database schema integration. *ACM Computing Survey, 18*(4), 323–364.

Bruce, T., & Hillman, D. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. Hillman & E. Westbrooks (Eds.), *Metadata in practice* (pp. 238–256). Chicago: ALA Editions.

Bryant, S., Forte, A., & Bruckman, A. (2005). Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. In K. Schmidt, M. Pendergast, M. Ackerman, & G. Mark (Eds.), *Proceedings of GROUP International Conference on Supporting Group Work* (pp. 11–20). New York: ACM Press.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley.

Emigh, W., & Herring, S. (2005). Collaborative authoring on the web: A genre analysis of online encyclopedias. In R. Sprague (Ed.), *Proceedings of the 38th Hawaii International Conference on System Sciences.* (p. 99a). Los Alamitos, CA: IEEE Computer Society Press.

Eppler, M. (2003). *Managing information quality: Increasing the value of information in knowledge-intensive products and processes*. Berlin, Germany: Springer-Verlag.

Fallis, D., & Frické, M. (2002). Indicators of accuracy of consumer health information on the Internet: A study of indicators relating to information for managing fever in children in the home. *Journal of the American Medical Informatics Association, 9*(1), 73–79.

Gasser, L., Sanderson, A., & Zdonik, S. (2007). *NSF IIS-GENI workshop report: First edition.* National Science Foundation. Washington, DC. Retrieved March 22, 2008, from https://apps.lis.uiuc.edu/wiki/download/attachments/10304/iis-geni-report-first-edition.pdf

Geertz, A. (1973). *Interpretation of cultures.* New York: Basic Books.

Halevy, A., Ashish, N., Bitton, D., Carey, M., Draper, D., Pollock, J., Rosenthal, A., & Sikka, V. (2005). Enterprise information integration: Successes, challenges and controversies. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data* (pp. 778–787). New York: ACM Press.

Hammwöhner, R. (2007). Interlingual aspects of Wikipedia's quality. In M. Robbert, R. O'Hare, M. Markus, & B. Klein (Eds.), *Proceedings of the 12th International Conference on Information Quality* (pp. 39–49). Cambridge, MA: MITIQ.

Hofstede, G. (1991). *Cultures and organizations: Software of the mind.* London: McGraw-Hill.

Hofstede, G., & McCrae, R. (2004). Personality and culture revisited: Linking traits and dimensions of culture. *Cross-Cultural Research, 38*(1), 52–88.

ITIM International. (1967–2009). *Geert Hofstede cultural dimensions*. Retrieved March 17, 2008, from http://www.geert-hofstede.com/hofstede_dimensions.php

Kittur, A., & Kraut, R. (2008). Harnessing the wisdom of crowds in Wikipedia: Quality through coordination. In *Proceedings of the ACM 2008 Conference on Computer-Supported Cooperative Work. CSCW '08.* New York: ACM Press.

Kittur, A., Suh, B., Chi, E., & Pendleton, B. (2007). He says, she says: Conflict and coordination in Wikipedia. In M. Rosson & D. Gilmore (Eds.), *Proceedings of CHI 2007* (pp. 453–462). New York: ACM Press.

Kroeber, A., & Kluckhorn, C. (1952). Culture: A critical review of concepts and definitions. *Harvard University Peabody Museum of American Archeology and Ethnology Papers 47,* 543-656.

Kuutti, K. (1991). Activity theory and its applications to information systems research and development. In H.-E. Nissen (Ed.), *Information systems research* (pp. 529–549). Amsterdam: Elsevier Science.

Leontiev, A. (1978). *Activity, consciousness, personality.* Englewood Cliffs, NJ: Prentice Hall.

Lih, A. (2004). Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of 5th International Symposium on Online Journalism*. Austin, TX. Retrieved March 17, 2009, from http://jmsc.hku.hk/faculty/alih/publications/utaustin-2004-wikipedia-rc2.pdf

Federal Bureau of Investigation. (2009). *N-DEx: Law enforcement national data exchange.* Retrieved March 17, 2009, from http://www.fbi.gov/hq/cjisd/ndex/ndex_home.htm

Pfeil, U., Zaphiris, P., & Ang, C. (2006). Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication, 12*(1). Retrieved March 17, 2009, from http://jcmc.indiana.edu/vol12/issue1/pfeil.html

Rask, M. (2007). The richness and reach of Wikinomics: Is the free web-based encyclopedia Wikipedia only for the rich countries? In *Proceedings of the Joint Conference of the International Society of Marketing Development and the Macromarketing Society*. Retrieved February 24, 2009, from the Social Science Research Network Web site: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=996158

Salton, G., & McGill, M. (1982). *Introduction to modern information retrieval.* New York: McGraw-Hill.

Star, S., Ruhleder, K. (1996). Steps toward an ecology of infrastructure: design and access for large information spaces. *Information Systems Research, 7*(1), 111 - 134.

Strong, D., Lee, Y., & Wang, R. (1997). Data quality in context. *Communications of the ACM, 40*(5), 103–110.

Stvilia, B. (2007, December). A model for ontology quality evaluation. *First Monday, 12*(12).

Stvilia, B., Gasser, L. (2008). An activity theoretic model for information quality change. *First Monday*, *13*(4).

Stvilia, B., Gasser, L., Twidale M. B., & Smith, L. C. (2007). A framework for information quality Assessment. *Journal of the American Society for Information Science and Technology, 58*(12), 1720–1733.

Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2005). Assessing information quality of a community-based encyclopedia. In F. Naumann, M. Gertz, & S. Mednick (Eds.), *Proceedings of the International Conference on Information Quality–ICIQ 2005* (pp. 442–454). Cambridge, MA: MITIQ.

Stvilia, B., Twidale, M., Smith, L. C., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology, 59*(6), 983–1001.

Swidler, A. (1986). Culture in action: symbols and strategies. American Sociological Review, *51*(2), 273-286.

TeraGrid. (2009). Retrieved March 17, 2009, from http://www.teragrid.org/

The Library of Congress (2008). *On the record: report of the library of congress working group on the future of bibliographic control.* Retrieved April 21, 2009, from http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf

Viegas, F., Wattenberg, M., Kriss, J., & van Ham, F. (2007). Talk before you type: Coordination in Wikipedia. In *Proceedings of HICSS 2007.* Retrieved February 19, 2007, from http://www.research.ibm.com/visual/papers/wikipedia_coordination_final.pdf

Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–35.

Wilkinson, D., & Huberman, B. (2007, April). Assessing the value of cooperation in Wikipedia. *First Monday*, *12*(4). Retrieved February 8, 2008, from http://journals.uic.edu/fm/article/view/1763/1643