Information Quality and Trustworthiness: A Topical State-of-the-Art Review

Jason R. C. Nurse, Syed S. Rahman, Sadie Creese, Michael Goldsmith, Koen Lamberts
University of Warwick, Coventry, CV4 7AL, UK
{j.nurse, s.s.rahman, s.creese, m.h.goldsmith, k.lamberts}@warwick.ac.uk

***Abstract*—The importance and value of *information* cannot be disputed. It is used as basis for menial and mission-critical tasks alike. In a society where information is so easily publicised and freely accessible, however, being able to assess information quality and trustworthiness is paramount. With appreciation of this fact, our paper seeks to navigate these two mature fields and define the latest state-of-the-art. The novelty of this work is found in the provision of an up-to-date review, a research survey which considers and links provenance, quality and trustworthiness, and a literature analysis that includes a first-look review at some of these aspects within the social-media domain. This factor-based review should provide an ideal grounding for future research that assesses interaction between these three topics, which may then also progress to associations with information assurance and security at large. To demonstrate how some of the factors might be considered, we also examine their application to a commonplace scenario.**

Keywords-information quality; trustworthiness; provenance; integrity; factors

I. INTRODUCTION

Information is at the centre of today's fast-paced world. People use it to make a range of decisions, from the very simple to the unbelievably complex. Regardless of the domain of interest, a general premise is: the more information the better. Apart from quantity however, information quality and trustworthiness are also crucial aspects, particularly in the human decision-making context. Adopting time-tested definitions, information quality considers the fitness of information for use [1], while information trustworthiness defines the perceived likelihood that a piece of information will preserve a user's trust in it, and encompasses characteristics such as the competence and predictability of the information source (adapted from [2]). A concept related to both of these topics, which is also worthy of note, is information provenance. Here, provenance refers to the source of information, including who produced it, what changes were made, amongst other aspects [3]. These three concepts, especially the first two, have occupied information-science discussions for many years, and justifiably so.

One way to consider information quality is as an enabler of information trustworthiness. Therefore, if information quality is low, a user is often likely to have less confidence that a piece of information will preserve their trust. The authors of [4] generally support this perspective, as they express that trustworthiness issues include quality and provenance issues. The link between information provenance and information trustworthiness can be seen in various articles ([4, 5]), including a study ([6]) which clearly mentions provenance as one of the core aspects that influence the trustworthiness of content. Furthermore, provenance may also aid in assessing information quality and thus determining the degree of trust that should be attributed to it [7].

With appreciation of the implicit importance and underlying relationship between these topics, in this paper we aim to provide a topical state-of-the-art review not yet addressed in the current literature. This review's novelty stems from three aspects. First, it conducts an up-to-date analysis of the continually progressing research fields of information provenance, quality and trustworthiness. Second, the review assesses all three fields and the links between fields, also considering both offline and online contexts. Finally, within the work, we specially include a first-look review at the social-media domain; the importance of this domain is clear, noting the significant amount of information content attributable to it online. The social-media domain covers services such as Twitter, Wikipedia, Yahoo! Answers and Facebook. This paper's review takes a factor-oriented approach, and therefore during our discussions seminal factors which influence information provenance, quality and trustworthiness are observed. This is beneficial for several reasons, but a prime one is that it consolidates research fields as they currently are, and establishes a core set of factors on which researchers can ground future work. For ease of reference, throughout the paper new factors/properties are italicised the first time they are encountered.

This paper is structured as follows. Section II examines information provenance as it relates to quality and trustworthiness. Next, Section III begins the core research and considers information quality and the factors/dimensions within it. After this, we move on to review the field of information trustworthiness in Section IV. Section V seeks to demonstrate the real-world application and interplay between some of these factors through the use of a commonplace wiki-based scenario, before we conclude the paper in Section VI.

II. INFORMATION PROVENANCE

To recap, provenance of information refers to the source of information, such as who produced it, its derivation history, what data was used to generate it, and also the trail

978-1-4244-9764-5/11/$26.00 ©2011 IEEE

of how the information passed between sources and how it has been altered ([3, 8]). Provenance is central to the fields of databases and workflow systems in particular. There are many factors/properties proposed in the literature for determining provenance, some explicitly and others implicitly designed to provide evidence of measures of information quality and trustworthiness. In this review, we adopt a factor-by-factor approach considering the small set of provenance factors found in the literature.

A central component of information provenance is the information source, i.e. who produced or changed the information. As a result, the characteristics of a source become useful clues in forming judgements on quality and trustworthiness. The first factor therefore is the actual *identity* of the source. Identity supplies a base for provenance, risk assessment and trustworthiness [5, 9]. Knowing the identity of the information source might be regarded as a crucial initial step to attributing a level of trust to any information received. Another important provenance factor also related to an information source is the *location* of the source [8]. In a growing number of situations today (especially those that are news-related), the location from which a source reports has an influence on the believability and trustworthiness of the information supplied [10]. It is typical, for example, to give greater credence to information received from an eyewitness about an incident as opposed to someone who is miles away.

In addition to examining characteristics of the information source, the attributes of the information itself are useful for provenance, quality and trustworthiness deliberations. *Freshness or timeliness* is one such factor grounded within provenance that determines the use of information [8, 11]. A receiver of information therefore needs to know when it was produced to judge how fresh or contemporaneous it is and assign a measure of quality or trustworthiness accordingly.

The next provenance factor seen in the literature is the *motivation* or reasons why the information was produced [8]. This adds value to quality and trustworthiness judgements as it supplies support for a source's actions, which decision-makers can then use to assess a final information object. As noted in [8], how the information event occurred and which instruments or programs were used, are useful provenance considerations as well. But these can be regarded as very context-dependent, as they are not always of great use.

Next we move on to assess the core topics of this paper: information quality and trustworthiness. Where provenance factors relate, we specially highlight this and thus identify relationships across the topics.

III. Data and Information Quality

Data and information quality have been of interest to researchers and practitioners for decades. As such, there has been a plethora of related studies, reports and publications. In this section we review the most significant and relevant of these articles. Similar to various articles in the literature (as supported by [12]), within this report data and information quality are regarded as synonymous unless otherwise stated.

Information quality can be defined as an assessment or measure of how fit an information object is for use. This notion of 'fitness for use' is central to several information quality discourses and is apparent in numerous research articles ([1, 12, 13, 14, 15]). A crucial question which surfaces in most quality literature is, what are the dimensions/factors that comprise information quality and thus lead to its final valuation. We address this question next in a largely chronological order, and specially concentrate on the most significant contributions, surveys and literature reviews/summaries. This narrowed focus appreciates paper-space limitations and enables us to gather the greatest number of core quality dimensions/factors from the smallest number of articles. The emphasis on core dimensions means that unless newly proposed factors are novel and widely applicable, they will be grouped with closely related established factors.

Undoubtedly, one of the most significant and popular contributions within the field of information-quality research is found in [1]. In this work, the authors conduct a comprehensive review of quality dimensions and develop a novel hierarchical framework for data quality. The framework defines four quality groups. Intrinsic quality is the first group and stresses the fact that in its own right, information possesses a level of quality. Dimensions within this group are *accuracy* (correct, reliable), *believability* (regarded as true and credible), *objectivity* (unbiased) and *reputation* (trusted in terms of source or content). The contextual quality group advances the discussion and emphasises that quality cannot be judged without assessing the context at hand. *Value-added* (provide advantages from use), *relevancy* (applicable and helpful), *timeliness* (age of the data is appropriate), *completeness* (sufficient breadth, depth and scope) and *appropriate amount of data* (appropriate volume of data available) are therefore identified within this area. The representational quality group focuses on data representation aspects and includes *interpretability* (appropriate language and units and data definitions are clear), *ease of understanding* (without ambiguity and easily comprehended), *representational consistency* (always presented in the same format and are compatible with previous data) and *concise representation* (compactly represented without being overwhelming) dimensions. Finally, the accessibility quality group regards quality in the light of the ease with which desired information can be obtained and restricted as necessary. As such, related dimensions are *accessibility* (available or easily and quickly retrievable) and *access security* (access to data can be restricted and hence kept secure).

Research progress since [1] has concentrated mainly on substantiating, specialising and extending a number of the quality dimensions mentioned above. In [13], where the authors define a semiotic-based framework for data quality, the aforementioned quality dimensions are generally maintained with accuracy, timeliness, reputation, accessibility and objectivity constituting examples of factors discussed. These findings are therefore seen to support the research in [1]. This research ([13]) did introduce two dimensions, i.e. the need for information to be

comprehensive and meaningful. Examining these factors critically however, in many ways they resemble specialisations and variations to completeness and value-added respectively (dictionary definitions for comprehensive and meaningful are assumed, as none were supplied). Considering the link to information provenance thus far we see the timeliness factor as being crucial to information quality.

The authors of [16] provide another useful resource of quality dimensions in the literature. Amongst their list of criteria are common factors such as relevance, understandability, amount of data, security and timeliness (age of information), but also new dimensions including *documentation* (amount and usefulness of documents with meta information), *verifiability* (degree and ease with which the information can be checked for correctness), availability (percentage of time an information source is 'up') and response time (amount of time until complete response reaches the user). Assessing these new dimensions, availability is seen to link to the previously identified accessibility and response time is similar to the general notion of timeliness. The other factors however are noteworthy. Reflecting briefly on verifiability, the provenance of information links strongly to this factor. Given that changes, authors of changes, time of changes and other lineage data has been recorded, verifying an information object is made much easier.

Reference [17] outlines dimensions akin to those presented above (e.g. interpretability, accessibility, reputation and value-added) but replace accuracy with free-of-error and introduce the need for *ease of manipulation* of information. The latter of these terms is new and especially concerned with the degree with which information can be easily manipulated and applied to different tasks. In another article assessing information quality, [18] outlines a generic quality framework built on quality dimensions. Once again however, numerous of the established dimensions reoccur (particularly comprehensiveness, clarity, correctness, accessibility and timeliness) with only a few new factors mentioned. The new dimensions include convenience, comprehensiveness, *interactivity* and *traceability*. Whereas convenience and comprehensiveness are variations on accessibility and completeness respectively, the last two factors are somewhat novel. Traceability in particular is also likely to be covered by provenance properties (e.g. identity and timing records). Here again we assume dictionary definitions for these terms.

Broadening the scope to information quality on the Web, [19] appreciates the lack of empirical studies in that field and therefore conducts such a study to identify factors which persons use in the judgement of information quality. Dimensions defined include *source* (origin of information is present), content (relevance, value added and *specificity*), *format* (structure), *presentation*, currency (timeliness), accuracy and speed of loading. The novelty of source comes from knowing the origin of information and characteristics of that origin (e.g. competence and affiliation); this is metadata that can be provided by the identity provenance factor. Specificity stresses the need of information to be not too general or abstract. Lastly, format and presentation are very similar and together look at structure, writing style and clarity. These are thought to overlap with and possibly even encapsulate established representational factors (e.g. representational consistency). Speed of loading, albeit particularly apt for the Web, is thought to be encompassed in response time, hence not a new factor.

In [12] and [20], researchers offer two comprehensive reviews of information quality literature in the Web domain. The first survey ([12]) conducted a thorough literature analysis and concluded by highlighting twenty of the most common information-quality dimensions in the field. Example factors include reliability (information is correct and reliable), *usability* (information is clear and easily used), efficiency (able to quickly meet information needs for task at hand) and navigation (easily found or linked to). Comparing these to the established dimensions, usability is novel but reliability links to accuracy, navigation arguably falls within accessibility, and efficiency is really speed of usability.

The second review ([20]) covers a longer period but authors choose to emphasise a smaller set of factors crucial to information quality online. Some of these are accessibility, timeliness, believability, appropriateness/relevance and source (the source of the information should be available). This second review overlaps the period assessed in the first and as such should be viewed more in light of its ability to verify and pick up on any factors missed in the initial survey. What is intriguing about the outcome of both these surveys and generally the research findings thus far, is that a majority of the dimensions identified link either directly or indirectly to those from [1]. Apart from reinforcing that initial work and extending application to the online space, this reality also hints at the possibility that the core dimensions of information quality might already be known at this point.

Recent work has continued the focus on information-quality dimensions, albeit only as a stepping stone towards the more novel goals of defining quality frameworks for specific applications or detailed quality-assessment approaches. Reference [21] exemplifies the former of these intentions, as they draw on over a decade of existing literature to outline an information-quality framework for e-learning systems. Their framework is adapted from [1] (and as such uses the hierarchical structure) but is updated to reflect recent literature and requirements in their application domain. Two of the new dimensions proposed are verifiability and response time, both of which we have discussed above.

As part of their charter towards a comprehensive information-quality framework, the authors of [22] build on established factors (e.g. accuracy, complexity, completeness and security) and also contribute new dimensions to judge quality. These include *cohesiveness* (extent to which an information object is concentrated on one topic), informativeness (the actual amount of informative content), complexity (how cognitively complex is it), consistency, volatility (the amount of time the information remains valid in the context of a particular activity), and authority (the level of reputation of an information object in a specified community). Although each of these possesses some degree

of novelty, apart from cohesiveness they are very similar to established quality dimensions. For example, complexity links to ease of understanding, volatility can generally compare to timeliness, and authority might be viewed as related to the original definition (in [1]) of reputation of content. Cohesiveness is unique because even though it may link to existing factors such as completeness, a complete piece of information is not necessarily cohesive.

Social-media websites have added significantly more complexity to the information science field. These websites break down standard barriers for publication and allow practically anyone to generate and publish online information content. Typical social media encompass services for weblogging, microblogging, photo sharing, social networking and wikis. What this new wave of services has meant is a significant increase in the quantity of online content, but also, large questions about its quality and trustworthiness. We therefore examine the literature in this topic as it relates to these areas.

In the social-media domain, it is apparent that several of the established information-quality factors maintain relevance. The authors of [23] support this point in their discussions of information quality in Wikipedia, undoubtedly the most popular wiki online. Common problem dimensions cited by Wikipedia users include accessibility (e.g. language barriers), accuracy (e.g. typos and conflicting reports of factual information), complexity (e.g. low readability) and verifiability (e.g. lack of references to original sources and lack of accessibility of original sources). In [24], researchers look generally at information quality and use the very simplistic notion of word count to assess quality in Wikipedia articles. Most importantly, their work does give some indication that article length may be a reasonable predictor as to whether or not an article will be featured (i.e. is of an acceptable quality). More research is needed to substantiate these claims and verify whether they can be generalised, particular to answer the question if length equals quality.

Reference [25] considers the weblog social-media domain and has the novel goal of seeking to prioritise information-quality factors/criteria by allocating priority coefficients. In order of importance, some of the factors they highlight are understandability, informativeness, representation, accuracy, completeness and timeliness. Other general criteria include cohesiveness, *maintainability*, *source popularity*, customer support and objectivity. Customer support is new to our review but it is regarded as too specific to the software characteristic of weblogs for our general information quality use. Information maintainability and source popularity however have wider applicability and thus are added to the factor list. In addition to studying a large range of factors, there has been work on assessing quality based on a smaller factor subset. Reference [26] exemplifies this, as it examines the information quality of weblog posts based on content depth and content breadth. Content depth builds on the completeness of each topic and the number of meaningful and useful words present in the weblog, while content breadth considers the topic variety (topic count, inter-topic distance – examining whether two topics in a post are related, and topic mergence – tracking down the hidden idea that happened on the combination of two post topics). Both of these, however, relate to completeness, as shown by [1].

With a large volume of the literature examined and arguably a majority of core information-quality factors determined, we conclude this review. Even though there is other recent research on quality (e.g. [15, 27, 28]), at first sight no new core concepts that might enhance this state-of-the-art review are apparent.

Before moving on to the next section, this report briefly assesses the notion of information integrity as it relates to information quality. This was intentionally postponed until this point as we thought that it would allow for a clearer discussion on two concepts that are often confused. Information integrity is most commonly defined as the representational faithfulness of information to the true state of the respective information object ([29, 30]). Representational faithfulness in this context speaks specifically to the accuracy/correctness, currency/timeliness, completeness and validity/authorisation of information. These attributes are mainstream, but as is to be expected there have been other suggestions, for example the work in [14] replaces currency and validity with consistency and existence as important dimensions. Generally comparing information integrity with information quality therefore, integrity can be regarded as a core part of quality which captures a cohesive subset of quality dimensions.

IV. Information Trustworthiness

Reference [6] provides an ideal starting point for our discussion on the most significant contributions, surveys and literature summaries in information-trustworthiness research. This work comprises of an exhaustive literature review into the factors that influence how end-users make decisions regarding trusting information. These factors include topic (trust in a resource is topic-dependent), *context and criticality* (context determines the criteria by which a user judges trustworthiness), *popularity* (widespread use of a resource tends to lead to more trust), *authority* (source identity and competence influence trust), *direct experience* (*reputation* leads to trust), *recommendation* (referrals from other users provide indirect reputation), *related resources* (relations to other entities which allow trust transferral), *bias* (biased sources may convey misleading information), *incentive* (information may be more believable if there is motivation for a resource to provide accurate information), *agreement* (*corroboration* influences trustworthiness), *age* (time of creation/lifespan of time-dependent data indicates its validity), *deception* (resources may have deceptive intentions), *specificity* (precise and specific content tends to engender more trust), *appearance* (user perception of a resource affects the trust of the content), *user expertise* (expert users may make better trust judgements on a resource's content), *limited resources* (absence of alternate resources may result in trusting imprecise information), *likelihood* (probability of content being correct, in light of everything known to the user), *recency* (content, associations and trust change with time) and *provenance*.

From a quick scan, one will notice that some factors recur from the discussion on information provenance. These include authority which links to identity, and age and recency which relate to freshness/timeliness of information. Assessing the description provided for authority, it is worth noting that this encompasses identity as well as *source competence*. This emphasises the need to have some knowledge on the expertise of an information source. Reverting to the general discourse in [6], these are all novel aspects from an information trustworthiness perspective.

In another comprehensive study, [2] synthesises existing work and constructs a useful framework for trust in information. The factors presented cover prediction (experience with the source), attribution (confirmation with multiple sources, i.e. agreement in [6]), reputation (reviews/references), *competence (of information)*, *positive intentions*, *ethics* (validity), *predictability* (persistent in both its presence and its contents, of information over time), social trust (recommendations), context (*relevance*), *bonding* (evocation of emotional response) and *propensity* (disposition to information). All of these, excluding prediction, attribution, reputation and recommendation, are new. Compared to the other articles, an interesting perspective held by [2] is that information has a level of competence itself. This therefore results in competence of source and competence of information (mentioned above). Within competence of information, the authors list core quality factors, namely, information accuracy, currency, coverage and believability. This highlights another link between quality and trustworthiness.

The authors of [31] also assess the topic of trust in information resources (e.g. websites) online. Their study is one of the most recent and in it they partition trust factors into three groups; external, internal and user's cognitive state. External factors assess external cues and cover aspects such as seals of approval, digital signatures proving authenticity of author and information, rankings, and recommendations from others. None of these however is particularly novel as seals of approval are a type of third-party recommendation, digital signatures prove identity/authority, recommendations were previously defined, and rankings are a variation of recommendations and/or demonstrated reputation. Internal factors define cues concentrated on the information itself and thus include reputation of source, *source motivation*, *accuracy*, objectivity (similar to bias), currency (or timeliness), coverage (*comprehensiveness*), presentation and format (similar to appearance), and citations (i.e. by whom has the information been cited; a variation on recommendation). Lastly, within the user's cognitive state (the end-user dimension), general factors are *disposition to trust*, *trust in general technology* and *risk propensity*. Source motivation is another factor which featured in our provenance review.

Complementary to previous research, there has also been work targeted at analysing trustworthiness within social-media domains. Work in [32] supplies one such study that concentrates on Wikipedia and identifying trustworthy articles. To assist in this task, the authors define several trustworthiness factors which assess aspects such as whether the article was written by expert and identifiable authors, if it is constantly visited and reviewed by authors, the presence of limited fragmentation of the contributions, the stability of the article, authors' use of a neutral point of view, a good presentation and format, and whether the article is well referenced. With the exception of the referencing factor, these all represent specialisations of established factors, spanning from competence of source to popularity, comprehensiveness, objectivity and presentation and format (inclusive of representation factors). The referencing of an article is novel as this concentrates on its *verifiability* by other users.

Assessing Wikipedia-article trustworthiness, [33] defines three main factors to examine, namely, reputation, performance and appearance. Performance is the most novel of these and considers the present conduct and current actions of an author, and/or user actions and responses towards the article content. This factor, however, strongly overlaps with several others including reputation (based on actions), recommendation (by other users) and authority (especially author competence).

From our literature survey, it is apparent that in some cases, trustworthiness is tied closely with credibility. Strictly speaking, for example, [31] outlines the factors above as factors affecting the 'trust/credibility' of information online. Other findings complement this and define credibility as a multifaceted concept with the two primary dimensions of trustworthiness and expertise [34]. With appreciation of this association and our goal to identify all core trustworthiness factors, we briefly review credibility aspects.

Reference [35] outlines a set of factors that affect information credibility online. These factors span source expertise/knowledge, credentials, *similarity to receiver beliefs/context* and goodwill. At the information/message level, factors relate to topic/content, *consistency/internal validity*, *plausibility of arguments* and *familiarity*. Finally, for the end-user, the factors include *motivation*, *beliefs*, issue relevance and prior knowledge on the issue. Considered in relation to trustworthiness (especially focused on information and information source), this is a generally well-accepted set of factors with the exception of similarity to receiver beliefs/context, consistency, plausibility of arguments and familiarity. (Credentials are thought of as a way to demonstrate authority/competence and goodwill as related to positive intentions.) Fast-forwarding to more recent work, [34] assesses credibility criteria as well and compares and discusses numerous key contribution articles. From these, the author notes four main factors that influence credibility judgements, i.e. authority, accuracy, comprehensiveness and objectivity [34]. All of these are very familiar concepts.

Specific domains within the social media have engaged in noteworthy research on credibility as well. In the weblogs domain, for example, [36] base a framework for weblog-credibility assessment on four factors. These encompass the blogger's (i.e. source's) expertise and the amount of offline identity disclosure (including, name and *geographic location*, credentials, and affiliations), the blogger's trustworthiness (including biases, beliefs and honesty), information quality (accuracy, completeness, and relevance), and appeals of a

personal nature (aesthetic appeal, literary appeal, curiosity trigger and personal connection). Reference [37] builds on that research and identifies a number of other credibility indicators targeted towards weblogs and posts. For posts, authors consider capitalisation, emoticons, shouting, spelling, post length and timeliness. As such, excessive use of emoticons, poor spelling, and constant shouting (using all capitals) are considered indicators of low-credibility information. At the weblog level, spam, comments, regularity and consistency are used by [37] to assess credibility. A weblog with regular posts and numerous third-party comments (i.e. where users judged that the blog was worth commenting on) is therefore regarded as more credible.

Although a few of the social-media factors/indicators above are domain-specific (e.g. comments and blogger consistency with weblogs), a majority have been encountered before. For example, beliefs and biases have been discussed, post length relates to completeness, and appearance and bonding address appeals of a personal nature. Two of the most interesting findings from the articles above are inclusion of source's geographic location (thereby creating another link across fields to the location provenance property) and the explicit analysis of literary aspects such as emoticons and shouting. The latter highlights another, more social media-targeted dimension within the appearance (presentation and format) factor reviewed previously.

Microblogging is another domain of interest slightly different to weblogs. Reference [38] provides one of the more topical and grounded works in this field, which studies the credibility of information on Twitter, a leading microblogging service. It identifies four types of features with which credibility might be judged. Message-based features include its length, presence of hashtags (keywords prefixed with #), whether or not the text contains question or exclamation marks and the number of positive/negative sentiment words in a message. User-based features assess registration age, number of followers, number of followees and the number of tweets the author has previously written. Topic-based features consider aspects such as the fraction of tweets that contain URLs, while propagation-based features include the depth of the re-tweet tree, or the number of initial tweets of a topic.

From this study, [38] concludes that credible (or newsworthy in their context) topics tend to include URLs and to have deep propagation trees. Furthermore, credible information is likely to have many re-posts, to originate at a few users in the network and to be propagated by authors with a vast number of previous tweets. In terms of our review, several of the features mentioned are captured by existing factors. In detail, verifiability addresses the inclusion of URLs, popularity of source links to number of followers, popularity of information covers re-posts, reputation relates to amount of useful previous tweets, and message length to comprehensiveness of content. Monitoring the use of question or exclamation marks and sentiment words can be seen to associate with assessing literary style, and thus the appearance factor. Returning to the use of punctuation marks, [38] note that tweets with question marks (or indeed other things such as smiling emoticons) are likely to be more related to non-credible information. This is an intriguing finding and should further research be found to substantiate it, it will undoubtedly act a key factor in judging credibility.

With our review of factors which influence information provenance, quality and trustworthiness complete, Table I summarises them. In the Type column, we identify whether factors are related to provenance, quality and trustworthiness with P, Q and T respectively. This also more clearly identifies similar factors across topics.

V. Application Scenario: Wikipedia

To demonstrate the real-world application and interplay between some of these factors, we briefly consider a Wikipedia (article) scenario. This type of scenario typically consists of numerous contributors (anonymous and otherwise), various edits/updates, and generally the sharing of large amounts of information, all towards the creation of a single Wikipedia article. As a result of how open and dynamic these environments are, a key concern always relates to the quality and trustworthiness of content provided, hence our interest in it here. For our discussion we focus on the Wikipedia article on Bottled Water [39]. This article defines bottled water and presents its effects, position in the marketplace and the regions of use. From a quality and trustworthiness perspective, the question therefore is, how does a reader decide that an article is of a suitable quality and should be trusted.

Based on our factor list in Table I, one of the initial ways in which a reader may assess information quality and trustworthiness is to determine who (*identity*) created/edited the article. Wikipedia provides an easy way to do this by allowing readers to freely view an article's history. A scaled-down snippet of the Bottled Water article history is below.

4: 16:34, 23 February 2011 ClueBotNG (contribs) (Reverting possible vandalism by <IP address #1> to version by Aspirex. False positive? Report it.)
3: 16:34, 23 February 2011 <IP address #1> (contribs) (=>Pakistan)
2: 11:48, 23 February 2011 Aspirex (contribs) (=>Bottled water ban in Bundanoon)
1: 11:28, 23 February 2011 Aspirex (contribs) (=>Effects of bottled water: merging 'bottled water phenomenon' article into this article)

The first point of note from this excerpt is that this history supplies a perfect example of information provenance and lineage. Consequently each line captures the *identity* of the associated source, the time the article was changed (from which one can infer *timeliness*), and in some situations (e.g. line 4) the *motivation* behind the change; the change itself is displayed on another wiki page as well. This in itself can serve as a basis for readers' initial quality and trust judgements. An interesting reality from the snippet above is that edits made by anonymous sources (with only IPs listed) are in themselves not blindly trusted by Wikipedia. As such, line 3 shows an anonymous edit but line 4 shows a quick deletion of that edit (by a bot program) citing possible vandalism. These provenance properties significantly aid

TABLE I. INFORMATION PROVENANCE, QUALITY AND TRUSTWORTHINESS FACTORS COMBINED

Factors	Type	Factors	Type
Access security	Q	Incentive	T
Accessibility (Availability, Convenience, Efficiency, Navigation)	Q	Interactivity	Q
Accuracy (Free-of-error, Reliability)	Q, T	Interpretability	Q
Appropriate amount of data	Q	Limited resources	T
Believability (Likelihood, Plausbility of arguments)	Q, T	Location of source (Geographic location)	P, T
Bonding	T	Maintainability	Q
Cohesiveness	Q	Objectivity (Bias)	Q, T
Completeness (Comprehensive, Content depth and breadth)	Q	Popularity	Q, T
Competence of information	Q, T	Positive intensions (Goodwill)	T
Consistency/Internal validity	T	Predictability	T
Context and criticality	T	Presentation and format (Appearance, Appeals of a personal nature, Representational consistency, Concise representation)	Q, T
Corroboration (Agreement)	T	Provenance	Q, T
Deception	T	Recommendation (Seals of approval, Rankings, Citations)	T
Documentation	Q	Related resources	T
Ease of manipulation	Q	Relevance	Q, T
Ease of understanding (Understandability, Complexity)	Q	Reputation (Direct experience, Prediction)	Q, T
End-user beliefs	T	Source motivation	P, T
End-user disposition to trust	T	Specificity	Q, T
End-user expertise	T	Similarity to receiver beliefs/context	T
End-user motivation	T	Timeliness/Freshness (Age, Recency, Volatility, Response time, Speed of loading)	P, Q, T
End-user propensity	T	Topic	T
End-user risk propensity	T	Traceability	Q
End-user trust in general technology	T	Usability	Q
Ethics	T	Value-added (Meaningful, Informativeness)	Q
Familiarity	T	Verifiability	Q, T
Identity (Source, Authority/Competence of source, Credentials, Digital signatures)	P, Q, T		

readers in making appropriate quality and trust judgements. Having determined the identity of the source responsible for an edit and seen their motivation, the next aspect readers may want to consider (particularly when a source is virtual) is the source's *reputation*. As reputation is typically based on direct experiences, in this scenario one question for readers would therefore be, whether there was any previous history of interactions with contributor Aspirex. Positive past interactions may lead to increases in perceived quality and trustworthiness of that source's changes, whereas negative interactions are likely to result in the opposite. As required, the 'contribs' link in the line records may be used to see a complete list of past contributions.

With some idea of identity of contributors, their reputation, and the motivation for changes/edits, readers may then examine the wiki article itself. Some of the aspects likely to be questioned here are the article's *believability*, *objectivity*, *relevance*, *traceability*, verifiability, *timeliness* and *ease of understanding*. This process generally consists of readers examining the article in terms of these factors. Taking a simple example, the Wikipedia Bottled water article states, "Consumption of water often is considered a healthier substitute for sodas[21]" [39]. Readers may therefore begin by assessing this information based on their knowledge and determine its believability, and consequently the quality and trust to be linked to that information snippet. While making that judgement, the traceability and verifiability factors become important considerations as well. This is because a reference (i.e. "[21]") is presented which highlights where the information was found and that may additionally be used to verify the claim made. A verifiable claim, especially one from a strong reference/source (with associated *competence* and/or reputation) is likely to increase perceived quality and trustworthiness of content. In this case the reference points to a Consumer Reports survey conducted by competent sources (top nutrition researchers) in 2006. This therefore suffices in terms of traceability and verifiability but possibility raises concerns over timeliness of information. For example, new research might have been done since 2006 that disproves this claim. Quality and trustworthiness may therefore be negatively affected in the eyes of readers.

When readers evaluate objectivity, the question is whether any bias may be present in the information or associated content. In addition to analysing the Wikipedia quote mentioned above therefore, it is also useful if readers check the related reference for bias. If the survey article was published by a company that sells bottled water for example, the objectivity and therefore quality and trustworthiness of that reference may come into question. Finally, readers may also consider ease of understanding in their judgements. Therefore, because the quote is simple and easily comprehended, it is likely to engender more feels of quality and trust in readers' minds than content which is ambiguous or confusing.

With appreciation of space limitations, we end our scenario discourse here. At a general level, however, all of the other factors from Table I may be applied in similar ways to a wide range of areas and information-based situations. Having given a practical example of how some of the factors might be applied to a real-world scenario and touched briefly on their interplay, the next section concludes this paper.

VI. Conclusion and Future Work

In this paper we have conducted a thorough review of the information-quality and trustworthiness fields (inclusive of information provenance) and highlighted core influential factors and properties. As such, this paper updates existing research by providing a topical review on these increasingly important topics. To demonstrate the practical application and interplay between some of these factors, we have considered a Wikipedia-based scenario in which we briefly discussed how factors might be used to aid users in forming judgements on the quality and trustworthiness of information articles. Considering the relationships and interplay between the factors identified, one avenue for future work is to examine exactly how the factors relate to each other. This may lead to a precedence model of factors that would be applicable and useful across all three domains. Another avenue would then be to expand further and assess associations with information assurance and security.

Acknowledgment

This work was conducted as a part of the TEASE project, a collaboration between the University of Warwick, HW Communications Ltd and Thales UK Research and Technology. The project is supported by the UK Technology Strategy Board Trusted Services Competition (www.innovateuk.org) and the Research Councils UK Digital Economy Programme (www.rcuk.ac.uk/digitaleconomy). Rahman is supported by a grant funded by the UK Engineering & Physical Sciences Research Council and Thales Research & Technology.

References

[1] R. Wang and D. Strong, "Beyond accuracy: What data quality means to data consumers," Journal of Management Information Systems, vol. 12, no. 4, pp. 5–34, 1996.

[2] K. Kelton, K. R. Fleischmann, and W. A. Wallace, "Trust in digital information," Journal of the American Society for Information Science and Technology, vol. 59, no. 3, pp. 363–374, 2008.

[3] O. Hartig, "Towards a data-centric notion of trust in the semantic web," in 2nd Workshop on Trust and Privacy on the Social and Semantic Web, 2010.

[4] E. Bertino and H.-S. Lim, "Assuring data trustworthiness - concepts and research challenges," in Secure Data Management, ser. Lecture Notes in Computer Science, W. Jonker and M. Petkovic, Eds., 2010, vol. 6358, pp. 1–12.

[5] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu, "An approach to evaluate data trustworthiness based on data provenance," in Fifth VLDB Workshop on Secure Data Management, ser. LNCS, W. Jonker and M. Petkovic, Eds. Springer, 2008, vol. 5159, pp. 82–98.

[6] Y. Gil and D. Artz, "Towards content trust of web resources," Journal of Web Semantics: Science, Services and Agents on the World Wide Web, vol. 5, no. 4, pp. 227–239, 2007.

[7] W. C. Tan, "Provenance in databases: Past, current, and future," IEEE Data Engineering Bulletin, vol. 30, no. 4, pp. 3–12, 2007.

[8] S. Ram and J. Liu, "A new perspective on semantics of data provenance," in 1st Workshop on the Role of Semantic Web in Provenance Management, 2009.

[9] S. Xu, R. Sandhu, and E. Bertino, "TIUPAM: A framework for trustworthiness-centric information sharing," in Third IFIP International Conference on Trust Management, ser. IFIP AICT, E. Ferrari, Ed., vol. 300, 2009, pp. 164–175.

[10] S. Toivonen and G. Denker, "The impact of context on the trustworthiness of communication: An ontological approach," in ISWC Workshop on Trust, Security, and Reputation on the Semantic Web, 2004.

[11] O. Hartig, "Provenance information in the web of data," in Linked Data on the Web Workshop at WWW, 2009.

[12] S.-A. Knight and J. Burn, "Developing a framework for assessing information quality on the world wide web," Informing Science Journal, vol. 8, pp. 160–172, 2005.

[13] G. Shanks and B. Corbitt, "Understanding data quality: Social and cultural aspects," in 10th Australasian Conference on Information Systems, 1999, pp. 785–797.

[14] M. Bovee, R. Srivastava, and B. Mak, "A conceptual framework and belief-function approach to assessing overall information quality," International Journal of Intelligent Systems, vol. 18, p. 5174, 2003.

[15] K. Keeton, P. Mehra, and J. Wilkes, "Do you know your IQ?: A research agenda for information quality in systems," SIGMETRICS Performance Evaluation Review, vol. 37, no. 3, pp. 26–31, 2010.

[16] F. Naumann and C. Rolker, "Assessment methods for information quality criteria," in 5th International Conference on Information Quality, 2000, pp. 148–162.

[17] L. Pipino, Y. Lee, and R. Wang, "Data quality assessment," Communications of the ACM, vol. 45, pp. 211–218, 2002.

[18] M. Eppler, M. Helfert, and U. Gasser, "Information quality: Organizational, technological, and legal perspectives," Studies in Communication Sciences, vol. 4, pp. 1–16, 2004.

[19] S. Rieh and N. Belkin, "Understanding judgment of information quality and cognitive authority in the WWW," in 61st ASIS Annual Meeting, 1998, pp. 279–289.

[20] M. Parker, V. Moleshe, R. De la Harpe, and G. Wills, "An evaluation of information quality frameworks for the World Wide Web," in 8th Annual Conference on WWW Applications, 2006.

[21] M. Alkhattabi, D. Neagu, and A. Cullen, "Information quality framework for e-learning systems," Knowledge Management & E-Learning: An International Journal, vol. 2, no. 4, pp. 340–362, 2010.

[22] B. Stvilia, L. Gasser, M. Twidale, and L. Smith, "A framework for information quality assessment," Journal of the American Society for Information Science and Technology, vol. 58, no. 12, pp. 1720–1733, 2007.

[23] B. Stvilia, M. Twidale, L. Gasser, and L. Smith, "Information quality discussions in wikipedia," Tech. Rep. ISRN UIUCLIS–2005/2+CSCW, 2005.

[24] J. Blumenstock, "Size matters: word count as a measure of quality on wikipedia," in 17th International Conference on World Wide Web, 2008, pp. 1095–1096.

[25] M. Kargar, A. Ramli, H. Ibrahim, and F. Azimzadeh, "Formulating priory of information quality criteria on the blog," World Applied Sciences Journal, vol. 4, no. 4, pp. 586–593, 2008.

[26] M. Chen and T. Ohta, "Using blog content depth and breadth to access and classify blogs," International Journal of Business and Information, vol. 5, no. 1, pp. 26–45, 2010.

[27] S. Madnick, R. Wang, Y. Lee, and H. Zhu, "Overview and framework for data and information quality research," Journal of Data and Information Quality, vol. 1, no. 1, 2009.

[28] O. Arazy and R. Kopak, "On the measurability of information quality," Journal of the American Society for Information Science and Technology, vol. 62, no. 1, pp. 89–99, 2011.

[29] IT Governance Institute (ITGI), Managing enterprise information integrity. Rolling Meadows, IL: ISACA, 2004.

[30] J. Boritz, "IS practitioners views on core concepts of information integrity," International Journal of Accounting Information Systems, vol. 6, pp. 260–279, 2005.

[31] A. J. Pickard, P. Gannon-Leary, and L. Coventry, "Trust in E: Users trust in information resources in the web environment," in ENTERprise Information Systems, ser. Communications in Computer and Information Science, J. E. Quintela Varajo, Ed., 2010, vol. 110, pp. 305–314.

[32] P. Dondio, S. Barrett, S. Weber, and J. Seigneur, "Extracting trust from domain analysis: A case study on the wikipedia project," in Autonomic and Trusted Computing, ser. Lecture Notes in Computer Science, L. Yang, H. Jin, J. Ma, and T. Ungerer, Eds., 2006, vol. 4158, pp. 362–373.

[33] S. Moturu and H. Liu, "Quantifying the trustworthiness of social media content," Distributed and Parallel Databases, pp. 1–22, 2010.

[34] M. Metzger, "Making sense of credibility on the web: Models for evaluating online information and recommendations for future research," Journal of the American Society for Information Science and Technology, vol. 58, no. 13, pp. 2078–2091, 2007.

[35] C. Wathen and J. Burkell, "Believe it or not: Factors influencing credibility on the web," Journal of the American Society for Information Science and Technology, vol. 53, no. 2, pp. 134–144, 2002.

[36] V. Rubin and E. Liddy, "Assessing credibility of weblogs," in AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 2006.

[37] W. Weerkamp and M. de Rijke, "Credibility improves topical blog post retrieval," in 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2008, pp. 923-931.

[38] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in 20th International Conference on World Wide Web, 2011.

[39] Wikipedia, "Bottled water," [Online]. Available: http://en.wikipedia.org/wiki/Bottled_water.