# Are Web Visibility and Data Quality Related Concepts?

Websites that provide a high level of data quality should logically be more visible for search engines. However, search engines base their results on methods such as Web hyperlink structure analysis and ignore data quality. The authors evaluate data quality in a group of 88 Web portals in three different domains and compare the results with the sites' positions in a visibility ranking. Looking at the relationship between data-quality level and visibility, their results show that data quality isn't a key factor in visibility.

**Angélica Caro**
*University of Bio Bio, Chile*

**Coral Calero and Mª Ángeles Moraga**
*University of Castilla-La Mancha, Spain*

Because the Internet has become such an important data source, it would be useful for people to know whether the data they obtain from a site has the appropriate level of data quality (DQ), especially when they're using that data for decision-making. However, search engine (SE) results don't provide details regarding DQ; they depend on website visibility – *visibility* being a website's ranking in SE results. Since users generally only look at the first few results displayed by the search,[1,2] results ranking is becoming increasingly significant. Thus, it would be interesting to determine whether DQ influences that visibility – that is, whether the first results in a ranking from an SE correspond to websites with a high level of DQ.

A website's visibility reflects its market reach and is particularly determined by how high SEs rank its pages (using methods such as Web hyperlink structure analysis for their indexation algorithms[3]) and popularity (www.thewebseye.com/website-visibility.htm). A website's popularity is a key factor in assuring its visibility – that is, how well known a site is by other websites (by means of links received and the popularity of the site that link to it) and by users (by means of the number of visits and the duration of these visits).

Therefore, despite the fact that DQ isn't directly used in websites for searches and visibility calculations, our premise is that if a site has high DQ then it should also have good visibility. Logically, if a website offers its users a high level of DQ, then the number of links from other sites would rise and its number of visitors would increase.


1089-7801/11/$26.00 © 2011 IEEE Published by the IEEE Computer Society

To test this premise, we evaluated the DQ in a group of 88 Web portals in three domains and compare these results with their position in a visibility ranking. We determine a website's visibility as a metric that combines site popularity (visitor analysis), number of pages from the site indexed by SEs, and number of links to the domain. We used two applications to develop this study: the Portal DQ Assessment (PoDQA; http://podqa.webportalquality.com) and the Know Your Visibility (KYV; http://kyv.webportalquality.com) tools.

## Determining Data Quality

A key aspect in determining a portal's success is the DQ level that it can offer to its users. In effect, the idea is that if a portal offers a level of DQ that satisfies its users' needs, it will obtain and maintain the users' preference.

DQ is often defined as "fitness for use" — that is, a data collection's ability to meet user requirements.[4,5] This definition emphasizes the current view of assessing DQ, which involves understanding it from the user's point of view as a key factor in evaluating it.[6] The ISO/IEC 25012 standard defines DQ as "the degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions." This definition encourages the vision of DQ as a multidimensional concept that depends on the context tackled.[4,7,8]

The magnitude of the data available on the Internet means that it's fundamental that a website be visible to its potential clients. We can calculate a site's visibility using data recovery, either using our own robot or the data available from SEs (such as Google, Yahoo, and MSN).

The first step in data retrieval is to delineate the set of candidate documents (contained in different types of websites) that are potential query answers. Owing to the Web's size, for most queries, the number of such candidates might be thousands or more. It's thus imperative to rank these documents in such a way that the first few results are likely to provide a satisfactory answer to the user's query.

## PDQM and the PoDQA Tool

The Portal Data Quality Model (PDQM) is a DQ model for Web portals that focuses on the data consumer perspective. Its definition was carried out in two phases. The first phase defines a theoretical model to obtain a set of DQ attributes that could be used to assess the DQ in portals, with the understanding that a DQ attribute is a measurable or observable characteristic. For our study, we used this phase to obtain a final set of 33 DQ attributes: attractiveness, accessibility accuracy, amount of data, applicability, availability, believability, completeness, concise representation, consistent representation, currency, documentation, duplicates, ease of operation, expiration, flexibility, interactivity, interpretability, novelty, objectivity, organization, relevancy, customer support, reliability, reputation, response time, security, specialization, timeliness, traceability, understandability, validity, and value added. (More details on the theoretical definition of PDQM are available elsewhere.[9])

The second phase consists of transforming the theoretical model into an operational one. In simple terms, this transformation consists of defining a structure with which to organize the DQ attributes and their relationships and then defining their measures. Taking into account the intrinsic subjectivity of the data consumer's perspective and the uncertainty inherent to quality perception,[10] we decided to use a probabilistic approach that uses Bayesian networks[11] for several reasons:

- to intuitively and explicitly represent the relationships between DQ attributes by connecting influencing factors to those that are influenced,
- to deal with the subjectivity and uncertainty implied in using probabilities,
- to use the network obtained to predict and estimate a Web portal's DQ, and
- to isolate factors responsible for low DQ.

This representation facilitates the model's comprehension, validation, evolution, and exploitation. We structured PDQM's attributes in a Bayesian network (BN) and organized them into four DQ categories (subnetworks): intrinsic, operational, contextual, and representational.

An initial step in the PDQM implementation consisted of working with the representational DQ subnetwork, which consists of eight attributes: interpretability, understandability, concise representation, consistent representation, amount of data, attractiveness, documentation, and organization. We implemented this subnetwork using the PoDQA tool.

We built the PoDQA tool so we could offer our model to Web portal data consumers. The objective was to provide users with a given portal's DQ level. In its first version (which we used in this work), the tool implements only PDQM's representational DQ category.

To evaluate a Web portal, the tool works exclusively with the public data in Web portals. It therefore downloads the portal pages and calculates their DQ level. The values of these indicators are transformed into a set of probabilities to be entered as evidence in the BN, and the BN is activated to calculate the level of representational DQ. PoDQA also stores the results of all the evaluations, thereby generating website rankings, which we organized according to the Web portal domain (such as universities, banks, or museums). (For more details about the tool and its implementation, see related work.[12])

The BN generates a set of values (or probabilities) for the representational DQ level. To help users, we transformed these probabilities into a descriptive category (high, high-medium, medium, medium-low, and low). We then send the description of the DQ level to the user.

## Visibility and KYV

To determine a website's visibility, the KYV tool uses cybermetric indicators that let us obtain a website's rank. To calculate some of these indicators, KYV uses the website's internal data, log files, HTML code, and so on. We calculate other indicators using public data.

To calculate the visibility KYV uses a four-phase method:

1. *Analyze website visibility*. The goal is to analyze a website's visitors and SEs, without taking its competitors into account.
2. *Identify competitors*. Two websites are competitors if their content or services belong to the same knowledge or business area and have a similar target public.
3. *Analyze competitor visibility*. After identifying competitors, KYV applies some of phase 1 to each competitor's sites, which lets us compare the best- and worst-ranked sites.
4. *Evaluate results and a plan of action*. After KYV obtains a general idea of a website's position, it's possible to develop a plan of action to define reasonable objectives.

The objective of the KYV tool is to show users the Web visibility of their website domains. This process can't take place in real time because we must ask the various SEs for different values several times to calculate the visibility indicators and prepare the rankings. In addition to providing visibility results, the application also offers activities that could be applied in the quest to improve that website's visibility.

The tool also stores statistics for each analyzed website so it can develop visibility rankings by topics. KYV computes a website's site visibility as a combination of the following parameters:

- *Indexed documents or pages*. Here, we determine the number of pages that the main SEs have indexed for a specific domain.
- *Domain links*. This value is a measurement of how well known the domain and its contents are on the Internet. These represent the number of recommendations that this domain receives as a link.
- *Alexa traffic rank* (popularity). Domain popularity attempts to determine how well known the domain is for users.

A more complete description of this method is available in related work.[13]

## DQ and Visibility

We used the PoDQA and KYV tools to obtain the representational DQ level and visibility of Web portals in three domains: Spanish universities, world museums, and Spanish computer science research groups. We then used the results we obtained to rank the portals. We finally compared these rankings, separated by domain, to determine whether any correlation exists.

### Measuring Web Portal DQ and Visibility

We first used PoDQA to evaluate the DQ of 73 Spanish university portals, obtaining an evaluation for 45 of them. (We couldn't evaluate the rest because we weren't able to obtain the HTML code PoDQA needed to evaluate the portals.) We then obtained the visibility ranking for those 45 sites. Table 1 shows the rankings.

We next looked at the DQ in a group of approximately 49 museums and obtained an evaluation for 26 of them. We then used KYV to obtain the visibility ranking of each of these museum portals. Table 2 shows the rankings obtained and the distance between them.

*Table 1. Data quality and visibility rankings for Spanish university portals.*

| Portal | DQ ranking | Visibility ranking | Partial visibility rankings | | | Distance* |
|---|---|---|---|---|---|---|
| | | | Site | Links | Popularity | |
| www.uma.es | 1 | 19 | 26 | 18 | 14 | 18 |
| www.uib.es | 2 | 16 | 13 | 10 | 25 | 14 |
| www.udl.es | 3 | 39 | 34 | 36 | 41 | 36 |
| www.us.es | 4 | 8 | 11 | 12 | 2 | **4** |
| www.ucm.es | 5 | 1 | 1 | 1 | 3 | **4** |
| www.uca.es | 6 | 37 | 38 | 39 | 26 | 31 |
| www.udc.es | 7 | 28 | 29 | 26 | 32 | 21 |
| www.unirioja.es | 8 | 12 | 4 | 29 | 1 | **4** |
| www.uv.es | 9 | 2 | 10 | 4 | 4 | 7 |
| www.unileon.es | 10 | 36 | 39 | 35 | 34 | 26 |
| www.uva.es | 11 | 21 | 22 | 19 | 15 | 10 |
| www.urv.es | 12 | 38 | 37 | 33 | 44 | 26 |
| www.upct.es | 13 | 44 | 44 | 45 | 42 | 31 |
| www.udg.es | 14 | 35 | 40 | 31 | 39 | 21 |
| www.ulpgc.es | 15 | 24 | 27 | 25 | 20 | 9 |
| www.uco.es | 16 | 31 | 35 | 30 | 29 | 15 |
| www.uah.es | 17 | 32 | 33 | 37 | 24 | 15 |
| www.ehu.es | 18 | 11 | 9 | 11 | 7 | 7 |
| www.uclm.es | 19 | 26 | 30 | 28 | 21 | 7 |
| www.usc.es | 20 | 25 | 25 | 20 | 33 | **5** |
| www.upm.es | 21 | 3 | 5 | 6 | 6 | 19 |
| www.upv.es | 22 | 5 | 3 | 7 | 11 | 18 |
| www.upf.edu | 23 | 29 | 31 | 24 | 43 | 5 |
| www.uc3m.es | 24 | 22 | 18 | 23 | 17 | **3** |
| www.ull.es | 25 | 30 | 24 | 32 | 28 | **4** |
| www.unizar.es | 26 | 6 | 6 | 5 | 13 | 21 |
| www.unav.es | 27 | 13 | 2 | 13 | 16 | 14 |
| www.unavarra.es | 28 | 43 | 41 | 43 | 40 | 15 |
| www.unia.es | 29 | 45 | 45 | 44 | 45 | 16 |
| www.um.es | 30 | 18 | 20 | 15 | 18 | 12 |
| www.urjc.es | 31 | 33 | 28 | 38 | 30 | **2** |
| www.unex.es | 32 | 27 | 23 | 27 | 31 | 5 |
| www.ujaen.es | 33 | 42 | 42 | 42 | 35 | 9 |
| www.ugr.es | 34 | 7 | 15 | 8 | 5 | 27 |
| www.usal.es | 35 | 17 | 19 | 14 | 19 | 18 |
| www.uniovi.es | 36 | 15 | 16 | 21 | 8 | 21 |
| www.ua.es | 37 | 9 | 7 | 9 | 10 | 28 |
| www.unican.es | 38 | 34 | 32 | 34 | 37 | 4 |
| www.umh.es | 39 | 40 | 36 | 40 | 36 | **1** |
| www.ub.edu | 40 | 4 | 8 | 3 | 12 | 36 |
| www.uvigo.es | 41 | 23 | 21 | 22 | 27 | 18 |
| www.uhu.es | 42 | 41 | 43 | 41 | 38 | **1** |
| www.uned.es | 43 | 14 | 14 | 16 | 9 | 29 |
| www.uji.es | 44 | 20 | 17 | 17 | 22 | 24 |
| www.upc.es | 45 | 10 | 12 | 2 | 23 | 35 |

*Distance between the data quality and visibility rankings. Teal numbers indicate the portals that are relatively close in both rankings.

*Table 2. Data quality and visibility rankings for museum portals.*

| Portal | DQ ranking | Visibility ranking | Partial visibility rankings | | | Distance* |
|---|---|---|---|---|---|---|
| | | | Site | Links | Popularity | |
| www.belvedere.at | 1 | 17 | 8 | 22 | 20 | 16 |
| www.museodelprado.es | 2 | 8 | 22 | 11 | 1 | 6 |
| www.museoreinasofia.es | 3 | 15 | 14 | 17 | 15 | 12 |
| www.serralves.pt | 4 | 21 | 11 | 24 | 18 | 17 |
| www.rijksmuseum.nl | 5 | 9 | 5 | 9 | 11 | **4** |
| www.moma.org | 6 | 1 | 2 | 1 | 4 | 5 |
| www.kunsthaus.ch | 7 | 16 | 9 | 18 | 19 | 9 |
| www.kunstmuseumbasel.ch | 8 | 20 | 13 | 19 | 23 | 12 |
| www.britishmuseum.org | 9 | 4 | 16 | 4 | 9 | 5 |
| www.tate.org.uk | 10 | 2 | 1 | 2 | 2 | 8 |
| www.vangoghmuseum.nl | 11 | 14 | 15 | 10 | 12 | **3** |
| www.uffizi.firenze.it | 12 | 19 | 26 | 15 | 26 | 7 |
| www.guggenheim-bilbao.es | 13 | 13 | 21 | 12 | 13 | **0** |
| www.gemeentemuseum.nl | 14 | 22 | 20 | 20 | 22 | 8 |
| www.modernamuseet.se | 15 | 18 | 17 | 16 | 17 | **3** |
| www.museothyssen.org | 16 | 11 | 7 | 14 | 10 | 5 |
| www.nationalgallery.org.uk | 17 | 7 | 6 | 7 | 5 | 10 |
| www.modernart.ie | 18 | 24 | 12 | 25 | 24 | 6 |
| www.guggenheim-venice.it | 19 | 6 | 18 | 5 | 7 | 13 |
| www.artic.edu | 20 | 5 | 4 | 6 | 6 | 15 |
| www.ambrosiana.it | 21 | 26 | 23 | 26 | 25 | 5 |
| www.musee-picasso.fr | 22 | 23 | 25 | 21 | 16 | **1** |
| www.louvre.fr | 23 | 3 | 3 | 3 | 3 | 20 |
| www.lacma.org | 24 | 10 | 19 | 8 | 8 | 14 |
| www.khm.at | 25 | 12 | 10 | 13 | 14 | 13 |
| www.museopicassomalaga.org | 26 | 25 | 24 | 23 | 21 | **1** |

*Distance between the data quality and visibility rankings. Teal numbers indicate the portals that are relatively close in both rankings.

Finally, we evaluated the DQ in a group of portals of Spanish computer science research groups. We obtained an evaluation for 16 of them. We then obtained the visibility for the same portals by using KYV. Table 3 shows the results.

## Comparing the Results

Based on the data in the three tables, it's surprising to note that there isn't a high correlation between the visibility and DQ ranking.

In Table 1 (university portals), only eight portals are relatively close in both rankings, a difference of between 0 and 4 (see the bold numbers in the last column). These portals represent 20 percent of the portals we evaluated, which is a far lower correlation than we had hoped to find. In fact, some portals show a great difference in their rankings — for example, the 36-point difference in www.udl.es and the 35-point difference in www.upc.es portals.

In Table 2 (museum portals), only six portals are relatively close in both rankings, a difference of between 0 and 4 (see the last column). These portals represent 17 percent of the total we evaluated. Once again, some portals show a great difference in their rankings — for instance, the www.louvre.fr portal shows a 20-point difference.

Finally, in Table 3 (computer science research group portals), the result is better, with 10 portals that are relatively close in both rankings (see the last column). These represent 63 percent of the total portals we evaluated. However, only two portals have the same place in the rankings (13 percent).

Table 3. Data quality and visibility rankings for Spanish computer science research group portals.

| Portal | DQ ranking | Visibility ranking | Partial visibility rankings | | | Distance* |
|---|---|---|---|---|---|---|
| | | | Site | Links | Popularity | |
| http://alarcos.inf-cr.uclm.es | 1 | 1 | 2 | 1 | 7 | **0** |
| http://lsi.ugr.es/~gedes | 2 | 10 | 7 | 7 | 9 | 8 |
| www.giro.infor.uva.es | 3 | 6 | 6 | 4 | 4 | **3** |
| www.ugr.es/~jcgranja | 4 | 16 | 15 | 14 | 15 | 12 |
| www.dei.inf.uc3m.es | 5 | 2 | 1 | 2 | 5 | **3** |
| http://grise.ls.fi.upm.es | 6 | 9 | 5 | 5 | 12 | **3** |
| http://gplsi.dlsi.ua.es/iwad | 7 | 7 | 13 | 6 | 2 | **0** |
| http://rosalia.dc.fi.udc.es/lbd | 8 | 14 | 9 | 8 | 16 | 6 |
| http://kybele.escet.urjc.es | 9 | 4 | 4 | 3 | 8 | 5 |
| www.onekin.org/ekinTeam | 10 | 12 | 12 | 13 | 10 | **2** |
| www.um.es/giisw | 11 | 8 | 8 | 15 | 6 | **3** |
| www.lcc.uma.es/~gisum/gisumppal-es.html | 12 | 11 | 10 | 10 | 13 | **1** |
| www.lsi.upc.es/~webgessi | 13 | 15 | 11 | 11 | 14 | **2** |
| www.dsic.upv.es/users/elp/gplis.html | 14 | 5 | 3 | 9 | 3 | 9 |
| www3.uji.es/~berlanga/ | 15 | 13 | 14 | 16 | 11 | **2** |
| www.lsi.us.es/is | 16 | 3 | 16 | 12 | 1 | 13 |

*Distance between the data quality and visibility rankings. Teal numbers indicate the portals that are relatively close in both rankings.

Based on these results, there doesn't appear to be a relationship between a Web portal's DQ and visibility. If we take the number of all the portals together, only 25 out of 88 (28 percent) show a relationship between the two concepts, and again, there doesn't seem to be a relationship.

Considering that PoDQA assesses the DQ from the user's perspective, we decided to compare the DQ ranking with the popularity ranking (columns 2 and 6 in the tables), hoping that there might be a higher correlation. However, the results are similar.

These observations led us to conclude that, contrary to our expectations, DQ doesn't have a significant influence on a Web portal's visibility. From our point of view, this conclusion could provide important advice for developers. Based on these results, they would hardly need to work on DQ to make their websites more visible. In fact, other factors seem more relevant:

- How many pages of the website are indexed by an SE?
- Which website is linked from other pages or websites to cover a domain of interest but not necessarily as a result of the site's quality?
- How necessary is it for a group of users or a community to access or use a specific portal? Does this influence a website's popularity (for example, a bank's clients)?

Furthermore, although SEs have become commercialized (on average, 70 percent of the first pages of search results in most SEs are spidered results as opposed to being sponsored or featured listings), it's still important to optimize a website for SE spidering (www.thewebseye.com/website-visibility.htm).

Nevertheless, because organizations and individuals increasingly value data, providing and receiving data with an appropriate level of quality is increasingly more relevant and necessary. Consequently, sooner rather than later, it will be necessary for these SEs to include factors associated with quality as a key factor in the elaboration of the result rankings. Users are increasingly more demanding in this respect and will give priority to SEs that deliver query results that are the most appropriate to their needs.

Our results show that the portals with the best DQ aren't necessarily the most visible. We believe this is because quality isn't directly included in any of the factors considered by SEs in website visibility. Therefore, Web designers

don't need to consider DQ to make a website more visible. However, DQ should be considered as an important factor when designing a website to provide users with the most appropriate data for their needs.

In spite of the results of our study, we intend to continue investigating the possible relationship between a Web portal's DQ and visibility. We're even looking at the possibility of establishing new parameters through which to measure visibility that will account for aspects of quality that have probably not yet been considered. We must bear in mind that this study only partially evaluated DQ, only considering a submodel of PDQM, the representational DQ.

In the future, we plan to study a new strategy that will help us determine the relationship between a portal's DQ and visibility. We'll attempt to determine the relationship between each of the representational DQ subcharacteristics (representation, understandability, and attractiveness) and the three parameters identified in KYV (indexed pages, domain links, and the Alexa traffic rank) to calculate the distance between the concepts.

**Acknowledgments**

This research is part of the project "Operationalization and Implementation of a DQ Model for Web Portals" number 093619 2/R supported by the University of Bio Bio, Chile, PEGASO (Processes for the Improvement of Global Software Development, TIN2009-13718-C02-01); Medusas (Design Evaluation and Improvement, Usability, Security and Maintenance of Software) — CDTI (Center for the Industrial Technological Development) and MICINN (Ministry of Science and Innovation) (IDI-20090557) — supported by MICINN and FEDER (European Regional Development Found); DQNET (Network for the promotion of data quality in enterprises information systems) (TIN2008-04951-E) supported by MEC (Ministry of Education and Science); and IVISCUS (Indicators for the Quality in Use Visualization of Software Systems) project (PAC08-00245991) supported by the JCCM (Castilla-La Mancha Community Council).

**References**

1. J. Bar-llan et al., "User Rankings of Search Engine Results," *J. Am. Soc. Information Science and Technology*, vol. 58, no. 9, 2007, pp. 1254–1266.
2. P. Ferragina and A. Gulli, "A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering," *Software-Practice & Experience*, vol. 38, no. 2, 2007, pp. 189–225.
3. M. Henzinger, "Search Technologies for the Internet," *Science*, vol. 317, no. 5837, 2007, pp. 468–471.
4. C. Cappiello, C. Francalanci, and B. Pernici, "Data Quality Assessment from the User's Perspective," *Proc. Int'l Workshop Information Quality in Information Systems*, ACM Press, 2004, pp. 68–73.
5. D. Strong, Y. Lee, and R. Wang, "Data Quality in Context," *Comm. ACM*, vol. 40, no. 5, 1997, pp. 103–110.
6. S.A. Knight and J.M. Burn, "Developing a Framework for Assessing Information Quality on the World Wide Web," *Informing Science J.*, vol. 8, 2005, pp. 159–172.
7. L. Pipino, Y. Lee, and R. Wang, "Data Quality Assessment," *Comm. ACM*, vol. 45, no. 4, 2002, pp. 211–218.
8. R. Wang and D. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *J. Management Information Systems*, vol. 12, no. 4, 1996, pp. 5–33.
9. A. Caro, C. Calero, and M. Piattini, "A Proposal for a set of Attributes relevant for Web Portal Data Quality," *Software Quality J.*, vol. 16, no. 4, 2008, pp. 513-542.
10. M. Eppler, *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes*, Springer, 2003.
11. G. Malak et al., "Modeling Web-Based Applications Quality: A Probabilistic Approach," *Proc. 7th Int'l Conf. Web Information Systems Engineering*, LNCS, Springer, 2006, pp. 398–404.
12. A. Caro, C. Calero, and M. Piattini, "Development Process of the Operational Version of PDQM," *Proc. 8th Int'l Conf. Web Information Systems Eng.*, LNCS, Springer, 2007, pp. 436–448.
13. J. Espadas, C. Calero, and M. Piattini "Web Site Visibility Evaluation," *J. Am. Soc. Information Science and Technology*, vol. 59, no. 11, 2008, pp. 1727–1742.

**Angélica Caro** is an associate professor at the University of Bio Bio, Chile. Her research interests include data quality, data quality measurement, and legacy systems. Caro has a PhD in computer science from the University of Castilla-La Mancha. Contact her at mcaro@ubiobio.cl.

**Coral Calero** is a full professor and a member of the Alarcos Research Group at the University of Castilla-La Mancha, Spain. Her research interests include software quality, software measurement, Web and portal quality, and software quality models. Calero has a PhD in computer science from the University of Castilla-La Mancha. Contact her at coral.calero@uclm.es.

**Mª Ángeles Moraga** is an associate professor and a member of the Alarcos Research Group at the University of Castilla-La Mancha, Spain. Her research interests include Web portals, software quality, measures, software components, and visualization. Moraga has a PhD in computer science from the University of Castilla-La Mancha. Contact her at mariaangeles.moraga@uclm.es.