CLEI 2007

XXXIII CONFERENCIA LATINOAMERICANA EN INFORMÁTICA

Eventos adicionales

- XV Congreso Internacional de Educación Superior en Computación **(CIESC 2007)**
- Latin American Networking Conference **(LANC 2007)** organizada por la ACM y el IFIP
- I Taller Latinoamericano de Informática para la Biodiversidad **(INBI 2007)**
- XIV Concurso Latinoamericano de Tesis de Maestría

San José, Costa Rica,
del 9 al 12 de octubre de 2007

www.clei2007.org

UNIVERSIDAD DE COSTA RICA · UNED 30 · TEC · UNA · Ulatina · ULACIT · UNESCO · CAMTIC


XXXIII Conferencia Latinoamericana en Informática
San José, Costa Rica / 9-12 octubre 2007
CLEI 2007
001.6
C748m
Conferencia Latinoamericana en Informática
(23: 9-12 oct. 2007 : San José, Costa Rica)
Memoria de la XXXIII Conferencia Latinoamericana
en Informática : 2007. - San José, C.R. : UNED,
PEM, Oficina de Audiovisuales, 2007.
1 Disco compacto
ISBN 978-9968-9678-9-1
1. Informatica. I. Titulo.
Producción:
PEM

CLEI 2007

XXXIII Conferencia Latinoamericana de Informática

www.clei2007.org

San José, Costa Rica
9-12 Octubre 2007

Libro de Resúmenes

Editores

Marcelo Jenkins Coronas, Universidad de Costa Rica
Marta Calderón Campos, Universidad de Costa Rica
Diamela López Caurell, Universidad de Costa Rica
Rodrigo Bartels González, Universidad de Costa Rica

Diseño de portada

Alonso Gamboa Valverde, Universidad Estatal a Distancia

Tabla de contenidos

PRESENTACIÓN

La CLEI 2007 fue la trigésimo tercera versión de la Conferencia Latinoamérica en Informática realizada por primera vez en 1974. La edición de este año revistió de la particularidad de ser la primera ocasión en que se realizó con sede en Costa Rica y en un país centroamericano. Fue organizada por la Universidad de Costa Rica (UCR), la Universidad Estatal a Distancia (UNED), el Instituto Tecnológico de Costa Rica (ITCR), la Universidad Nacional de Costa Rica (UNA), la Universidad Latinoamericana de Ciencia y Tecnología (ULACIT) y la Universidad Latina de Costa Rica (ULATINA) con el apoyo de la Cámara de Tecnologías de Información y Comunicaciones (CAMTIC) y la UNESCO.

La CLEI 2007 renovó el objetivo de ser el foro latinoamericano más importante en el que científicos, investigadores, profesionales y estudiantes intercambian ideas y comparten resultados de investigaciones en Computación e Informática. Como parte de esta conferencia, se realizaron simultáneamente 5 eventos:

- XXXIII Conferencia Latinoamericana en Informática **CLEI 2007.**
- XV Congreso Internacional de Educación Superior en Computación **CIESC 2007.**
- Latin American Networking Conference **LANC 2007** organizada por la ACM y el IFIP.
- I Taller Latinoamericano de Informática para la Biodiversidad **INBI 2007.**
- XIV Concurso Latinoamericano de Tesis de Maestría

El evento comprendió diversas actividades, incluidas sesiones técnicas de presentación de trabajos, foros y paneles de discusión, así como charlas magistrales de personalidades destacadas en el campo.

Este año se recibieron en total 575 trabajos provenientes de más de 30 países y de más de 235 universidades y organizaciones. Fueron aceptados 159 trabajos para CLEI, 21 para CIESC, 17 para LANC y 20 para INBI, que junto con los 3 trabajos ganadores del Concurso de Tesis alcanzaron 220 trabajos presentados. En este libro se incluyen los resúmenes de todos los trabajos presentados como artículos y como posters durante los cuatro días del evento, así como los resúmenes de las conferencias magistrales y semi-magistrales y los tutoriales. El texto completo se encuentra en la memoria en el CD-ROM.

Queremos expresar un agradecimiento muy especial para todas aquellas personas que voluntariamente participaron en el proceso de revisión de los trabajos recibidos y que con ello contribuyeron significativamente a garantizar la alta calidad del evento.

Dr. Marcelo Jenkins
Co-presidente, Comité del Programa

Dr. Manuel E. Bermúdez
Co-presidente, Comité del Programa

COMITÉ ORGANIZADOR CLEI 2007

CO-PRESIDENTES COMITÉ ORGANIZADOR

Gabriela Marín, Universidad de Costa Rica

Marcelo Jenkins, Universidad de Costa Rica

COMITÉ ORGANIZADOR

Lidia Arévalo, Universidad de Costa Rica (Coordinadora comisión de logística)

Gabriela Barrantes, Universidad de Costa Rica (Coordinadora comisión página web)

Marta Calderón, Universidad de Costa Rica (Coordinadora comisión de divulgación)

Sanders Pacheco, Universidad de Costa Rica (Coordinador comisión de eventos sociales)

Nuria Rodríguez, UNED, Costa Rica

Yarima Sandoval, Instituto Tecnológico de Costa Rica

Yenory Rojas, ULATINA, Costa Rica

José A. Sánchez, Universidad Nacional, Costa Rica

Wilberth Molina, ULACIT, Costa Rica

Otto Rivera Valle, CAMTIC

PONENCIAS CLEI

Comité del Programa CLEI 2007

Manuel E. Bermúdez, U. of Florida at Gainesville manuel@cise.ufl.edu (Co-Presidente)
Marcelo Jenkins, Universidad de Costa Rica mjenkins@ecci.ucr.ac.cr (Co-Presidente)
Aldo Vecchietti, Centro Regional de Investigación y Desarrollo, Argentina
Georgina Stegmayer, Centro Regional de Investigación y Desarrollo, Argentina
Omar CHIOTTI, Centro Regional de Investigación y Desarrollo, Argentina
Hugo D. Scolnik, Universidad de Buenos Aires, Argentina
Irene Loiseau, Universidad de Buenos Aires, Argentina
Alicia Diaz, Universidad Nacional de La Plata, Argentina
Claudia Pons, Universidad Nacional de La Plata, Argentina
Marcelo Naiouf, Universidad Nacional de La Plata, Argentina
Gabriel Baum, Universidad Nacional de La Plata, Argentina
Nazareno M. Aguirre, Universidad Nacional de Río Cuarto, Argentina
Gabriela P. Henning, Universidad Nacional del Litoral, Argentina
Claudio Delrieux, Universidad Nacional del Sur, Argentina
Guillermo R. Simari, Universidad Nacional del Sur, Argentina
Ignacio Ponzoni, Universidad Nacional del Sur, Argentina
Pablo Villarreal, Universidad Tecnológica Nacional, Santa Fe, Argentina
Maria Rosa Galli, Centro Regional de Investigación y Desarrollo, Argentina
Ma. Laura Caliusco, Universidad Tecnológica Nacional, Argentina
Silvia Castro, Universidad Nacional del Sur, Argentina
Pablo Fillottrani, Universidad Nacional del Sur, Argentina
Mauricio Ayala-Rincon, Universidade de Brasilia, Brasil
Alex J. Cuadros Vargas, Universidade de Sao Paulo, Brasil
JC Maldonado, Universidade de Sáo Paulo, Brasil
Luis Rivera, Universidade Estadual do Norte Fluminense, Brasil
Miguel Jonathan, Universidade Federal do Rio de Janeiro, Brasil
Vanessa Braganholo Murta, Universidade Federal do Rio de Janeiro, Brasil
Daltro Jose Nunes, Universidade Federal do Rio Grande do Sul, Brasil
Luciano Paschoal Gaspary, Universidade Federal do Rio Grande do Sul, Brasil
Juliana Herbert, Unisinos, Brasil
Sergio Lifschitz, Pontificia Universidade Católica do Rio de Janeiro, Brasil
Giovani Librelotto, UNIFRA, Brasil
Yadran Eterovic, Pontificia Universidad Católica, Chile
Jorge Alfaro, Universidad Católica del Norte, Chile
Claudio Meneses, Universidad Católica del Norte, Chile
Gilberto Gutierrez, Universidad de Chile, Chile
Juan Alvarez, Universidad de Chile, Chile
Luis Guerrero, Universidad de Chile, Chile
Sergio Ochoa, Universidad de Chile, Chile
John Atkinson, Universidad de Concepción, Chile
Gonzalo Acuña Leiva, Universidad de Santiago de Chile, Chile
Hernán Astudillo, Universidad Técnica Federico Santa María, Chile
Marcello Visconti, Universidad Técnica Federico Santa María, Chile
Carlos Castro, Universidad Técnica Federico Santa María, Chile
Eric Monfroy, Universidad Técnica Federico Santa María, Chile
Carlos Pon, Universidad Católica del Norte, Chile
Federico Meza, Universidad de Talca, Chile
Homero Latorre, UTEM, Chile
Federico Meza, Universidad de Talca, Chile
Rodolfo Villarroel Acevedo, Universidad Católica del Maule, Chile
Camilo Rueda, Pontificia Universidad Javeriana, Colombia
Enrique González Guerrero, Pontificia Universidad Javeriana, Colombia

Fernando de la Rosa, Universidad de los Andes, Colombia
Jorge Villalobos, Universidad de los Andes, Colombia
Jose Tiberio Hernández, Universidad de los Andes, Colombia
Pablo Figueroa, Universidad de los Andes, Colombia
Rafael Armando García Gómez, Universidad de los Andes, Colombia
Rodrigo Cardoso, Universidad de los Andes, Colombia
Silvia Takahashi, Universidad de los Andes, Colombia
Alberto Restrepo, Universidad EAFIT, Colombia
Francisco José Correa Zabala, Universidad EAFIT, Colombia
Eduardo Carrillo, Universidad Autónoma de Bucaramanga, Colombia
Edwin Montoya, Universidad EAFIT, Colombia
Rafael David Rincón Bermúdez, Universidad EAFIT, Colombia
Cesar Garita, Instituto Tecnológico de Costa Rica, Costa Rica
Francisco Torres, Instituto Tecnológico de Costa Rica, Costa Rica
Pablo Alvarado, Instituto Tecnológico de Costa Rica, Costa Rica
Oscar López, Instituto Tecnológico de Costa Rica, Costa Rica
José Castro, Instituto Tecnológico de Costa Rica, Costa Rica
Carlos Loria, Universidad de Costa Rica, Costa Rica
Daniel Cañas, Intnet, Costa Rica, Costa Rica
Alvaro de la Ossa, Universidad de Costa Rica, Costa Rica
Francisco Mata, Universidad de Costa Rica, Costa Rica
Gabriela Barrantes, Universidad de Costa Rica, Costa Rica
Gabriela Marín, Universidad de Costa Rica, Costa Rica
Juan José Vargas, Universidad de Costa Rica, Costa Rica
Manuel Arce, Universidad de Costa Rica, Costa Rica
Oldemar Rodríguez, Universidad de Costa Rica, Costa Rica
Vladimir Lara, Universidad de Costa Rica, Costa Rica
Elzbieta Malinowski, Universidad de Costa Rica, Costa Rica
Carlos Vargas, Universidad de Costa Rica, Costa Rica
Luordes García, Universidad Central Marta Abreu de las Villas, Cuba
Alejandro Rosete, Instituto Superior Politécnico José Antonio Echeverría, Cuba
Martha Dunia Delgado, Instituto Superior Politécnico José Antonio Echeverría, Cuba
Francisco A. Cano, Universidad de las Ciencias Informáticas, Cuba
Gustavo Chafla, Datadec, Ecuador
Hugo Banda, Escuela Politécnica Nacional, Ecuador
Carmen Fernandez, Universidad Complutense de Madrid, España
Eusebi Calle, Universidad de Girona, España
Ramon Fabregat, Universidad de Girona, España
Ramon Puigjaner, Universidad de las Islas Baleares, España
Catalina Llado, Universidad de las Islas Baleares, España
Carlos Travieso González, Universidad de las Palmas en Gran Canaria, España
Jesús Alonso, Universidad de las Palmas en Gran Canaria, España
Miguel Ángel Ferrer Ballester, Universidad de las Palmas en Gran Canaria, España
Francisco-Jose Quiles Flor, Universidad Politécnica de Castilla La Mancha, España
Steven Wilmont, Universidad Politécnica de Cataluna, España
Javier Segovia, Universidad Politécnica de Madrid, España
Ana Pont, Universidad Politécnica de Valencia, España
Alberto Valderruten, UNIVERSIDADE DA CORUÑA, España
Oscar Hernández, Instituto Tecnológico de Monterrey, Querétaro, México
Begoña Albizuri, ITAM, México
Marcelo Mejia, ITAM, México
Guillermo Rodríguez, ITESM, México
Jose Incera, ITAM, México
Clifton E. Clunie B, Universidad Tecnólogica de Panamá, Panamá
Luca Cernuzzi, Universidad Católica, Paraguay
Benjamín Barán, Universidad Nacional de Asunción, Paraguay
Glen Rodriguez, Perú, Perú
Ernesto Cuadros-Vargas, Universidad Católica San Pablo, Perú
Sofia Alvarez, Universidad San Ignacio de Loyola, Perú
Nora Bertha La Serna, Universidad Nacional Mayor de San Marcos, Perú

Fernando Machado, Universidad Católica del Uruguay, Uruguay
Juan José Moreno, Universidad Católica del Uruguay, Uruguay
Héctor Cancela, Universidad de la República, Uruguay
Alberto Pardo, Universidad de la República, Uruguay
Alejandro Gutiérrez, Universidad de la República, Uruguay
Alfredo Viola, Universidad de la República, Uruguay
Alvaro Tasistro, Universidad ORT, Uruguay
Cristina Cornes, Universidad de la República, Uruguay
Dina Wonsever, Universidad de la República, Uruguay
Eduardo Grampán, Universidad de la República, Uruguay
Franco Robledo, Universidad de la República, Uruguay
Javier Couto, Universidad de la República, Uruguay
Juan José Cabezas, Universidad de la República, Uruguay
María E. Urquhart, Universidad de la República, Uruguay
Raúl Ruggia, Universidad de la República, Uruguay
Regina Motz, Universidad de la República, Uruguay
Sylvia da Rosa, Universidad de la República, Uruguay
Inés Kereki, Universidad Ort, Uruguay
Nora Szasz, Universidad Ort, Uruguay
Alberto J. Cañas, Institute for Human & Machine Cognition, USA
Daniel G. Chavarría, Pacific Northwest National Laboratory, USA
Ignacio Solis, Palo Alto Research Center, USA
Andrea Lobo, Rowan University, USA
Francisco Azuola, Temple University, USA
Joseph Berrios, University of Central Florida, USA
Avelino Gonzalez, University of Central Florida, USA
Manuel Núñez, University of Connecticut, USA
Ernst Leiss, University of Houston, USA
Eduardo B. Fernandez, Florida Atlantic University, USA
Lourdes Maritza Ortiz Sosa, Universidad Catolica Andres Bello, Venezuela
Wilmer Pereira, Universidad Catolica Andres Bello, Venezuela
Maria Elena Villapol, Universidad Central de Venezuela, Venezuela
Nancy Zambrano, Universidad Central de Venezuela, Venezuela
Nora Montaño, Universidad Central de Venezuela, Venezuela
Flor Narciso, Universidad de los Andes, Venezuela, Venezuela
Isabel Besembel, Universidad de los Andes, Venezuela
Jonás Montilva, Universidad de los Andes, Venezuela
Judith Barrios, Universidad de los Andes, Venezuela
Wladimir Rodríguez, Universidad de los Andes, Venezuela
Mariela Curiel, Universidad Simon Bolivar, Venezuela
Esther Vidal, Universidad Simon Bolivar, Venezuela
Yudith Cardinale, Universidad Simon Bolivar, Venezuela

Assessment of Web Portal Data Quality using Bayesian networks

Angélica Caro
Department of Computer Science and Information Technologies,
University of Bio Bio,
Chillán, Chile
mcaro@ubiobio.cl

and

Coral Calero, Mario Piattini
Alarcos Research Group. Information Systems and Technologies Department
UCLM-SOLUZIONA Research and Development Institute.
University of Castilla-La Mancha
Ciudad Real, Spain
{Coral.Calero,Mario.Piattini}@uclm.es

Abstract

The Internet has established itself as an environment for data administration, exchange and publication. A great variety of Web applications have appeared to support this phenomenon. Amongst these applications, Web portals are becoming an important means by which to access data. Many people need to use these applications to obtain data for different purposes: to do their work, to make decisions, for their business, for entertainment, etc. So, they need to assess the quality of data to ensure that what they have obtained is suitable for their needs.

In recent years, several research projects have been conducted on the topic of Web Data Quality. There is, however, still a lack of specific proposals for data quality in Web portals which take the data consumer's point of view into account. In this paper, we will introduce a data quality model for Web portals (PDQM). This model focuses on the point of view of data consumers and uses a probabilistic approach (based on Bayesian networks) for data quality evaluation.

Keywords: Data model, Data quality, Bayesian networks, Data quality assessment, Web portals.

Resumen

Hoy en día, se considera que Internet es un ambiente para la administración, intercambio y publicación de datos y para soportar este fenómeno ha surgido una gran variedad de aplicaciones Web. Entre ellas, los portales Web se han posicionado como un importante medio para el acceso a datos. Diariamente, muchas personas hacen uso de portales para obtener datos con distintos propósitos: desarrollar su trabajo, tomar decisiones, hacer negocios, por entretención, etc. Por lo que, ellos necesitarían poder evaluar si esos datos apropiados para el uso que desean darle. Es decir, necesitarían evaluar la calidad de los datos obtenidos.

En los últimos años, se han desarrollado diversos proyectos de investigación en el área de calidad de datos Web. Sin embargo, aún existe la necesidad de propuestas de modelos de calidad de datos específicas para portales Web, y que, además, consideren el punto de vista de los consumidores de datos en la evaluación de la calidad de los datos. En este artículo, presentamos un modelo de evaluación de la calidad de datos para portales Web (PDQM), que se centra en el punto de vista de los consumidores de datos y que emplea un enfoque probabilístico (basado de redes Bayesianas) para la evaluación.

Palabras clave: Modelo de datos, Calidad de datos, Redes Bayesianas, Evaluación de Calidad de datos, Portales Web.

1 Introduction

Web portal is a site that aggregates information from multiple sources on the World Wide Web and organizes this material in an easy user-friendly manner [33]. Over the past decade, the number of organizations which own Web portals has grown dramatically. They have established portals to complement, substitute or widen existing services to their clients [34]. Many people use data obtained from Web portals to carry out their work, as well as to make decisions. Users or data consumers of Web portals thus need to ensure that the data obtained from the portal are appropriate for the use they need.

In the relevant literature, the concept of Data Quality is often defined as "fitness for use", i.e., the ability of a data collection to meet users' requirements [6, 30]. Another point to remember is that the terms "data" and "information" are often used as synonyms [31] (as will be done in this work).

Research on DQ began in the context of information systems [18, 30] and has been extended to contexts such as cooperative systems, data warehouses or e-commerce [1, 4, 15, 21, 36], amongst others. Due to the particular characteristics of Web applications, the research community has begun to deal with the subject of DQ on the Web [12].

It is the case though, that there are no works on DQ that address the particular context of Web portals, in spite of the fact that some works do indeed highlight DQ as a relevant factor concerning the quality of a portal [23, 34]. On the other hand, a few works in the DQ area, like [5, 6, 11, 14, 32], have looked the DQ from the data consumers' perspective. This perspective differs from the data producers' and data custodians' perspectives in two important aspects [5]: (1) data consumers have no control over the quality of available data and (2) the aim of consumers is to find data that match their personal needs, rather than provide data that meet the needs of others.

Considering this shortcoming in the areas of DQ and Web portals, the aim of our research is to achieve the development of a Data Quality Model for Web portals that focuses on the data consumer's perspective. In the first part of our work, a theoretical model, named PDQM(t), was defined [7]. This model is composed of 33 DQ attributes that can be used to assess DQ in Web portals. In the second part, now in progress, our aim is to convert this theoretical model into an operative one. In simple terms, this conversion consists of the definition of a structure for organizing DQ attributes, along with the association of measures and criteria for these. Taking into account the intrinsic subjectivity of the data consumer's perspective and the uncertainty inherent to quality perception [9], we have decided to use a probabilistic approach (based on Bayesian networks and fuzzy logic), proposed in [20], to convert PDQM(t) into an operational model.

In this paper we will show the process developed for obtaining PDQM. The rest of the paper is organized as follows. Section 2 gives a brief discussion of DQ in the Web context. In Section 3 we describe the PDQM development process. The first part of this process and its results are summarized in Section 4. The approach used, the phases composing the second part of this process and their results are described in Section 5. Finally, Section 6 shows our conclusions.

2 Data Quality on the Web

Research on DQ began in the context of information systems [18, 30] and it has been extended to contexts such as cooperative systems, data warehouses or e-commerce, amongst others. Due to the particular characteristics of Web applications and their differences from traditional information systems, the research community started to deal with the subject of DQ on the Web [12]. In fact, the nature of the Web has forced us to pay attention to a series of typical issues of this context that can affect or influence DQ. The identification of these issues and others justifies the creating of specific proposals to deal with DQ on the Web.

Consequently, in the last few years frameworks and models to deal with DQ in different domains in the Web context have been proposed (among them [1, 10, 13, 15, 16, 22, 24, 25]). We can make two important observations concerning these works. Firstly, the frameworks proposed tackle different domains on the Web. This reasserts the idea that DQ needs to be assessed within the context of data generation [17]. Secondly, for Web portals we have not found specific DQ frameworks/ models.

During the past decade, an increasing number of organizations have established Web portals to complement, substitute or widen existing services to their clients. In general, portals provide users with access to different data sources (providers) [19], as well as to on-line information and information-related services [34].

In the same way, the amount of people that access these applications grows every day. They use these applications for different purposes, from business to education or entertainment. In each case, people need to perform operations related to data and they need data that are well-fitted to the use they wish to put them to. For example, if the purpose is to obtain the cinema billboard to find out the schedule of the movies, users will hope to receive the data that is suitable for programming what movie to watch and at what time, according to their plans. So they need data to be valid, correct, believable, accessible, etc; that is, they need data with quality.

DQ is a critical factor for the success of Web applications [28] nowadays. Web portals owners must know the DQ needs of data consumers if they are to ensure their loyalty. These consumers need to be assured that the data offered by a portal are appropriate for a given requirement. The challenge of our research, then, is to develop a DQ model for Web portals that will meet these needs.

3 A Portal Data Quality Model (PDQM)

The development of PDQM was organized into two parts, as shown in Figure 1. The first part corresponds to the theoretical definition of PDQM and meets the key aspects that represent the data consumer's perspective and the main characteristics of Web portals. It is composed of four phases. During the first phase, we have recompiled the Web DQ attributes from the literature which we believe should be applicable to Web portals. In the second phase we have built a matrix for the classification of the DQ attributes obtained in the previous phase. In the third phase we have used the matrix obtained to analyse the applicability of each Web DQ attribute to a Web portal. Finally, the fourth phase consisted in the validation of the preliminary model. The result of this part was used as an input to the second one.

The second part consists of the transformation of the PDQM(t) into an operational model. To do this, we have decided to use a probabilistic approach (discussed in section V). This second part is made up of four phases. During the first phase, we defined the criterion for organizing the DQ attributes of PDQM. In the second phase, we generated the graphical structure of PDQM (corresponding to the approach used). In the third phase, we prepared our model for its subsequent use in an evaluation process. Finally, the fourth phase corresponds to the validation of the model.

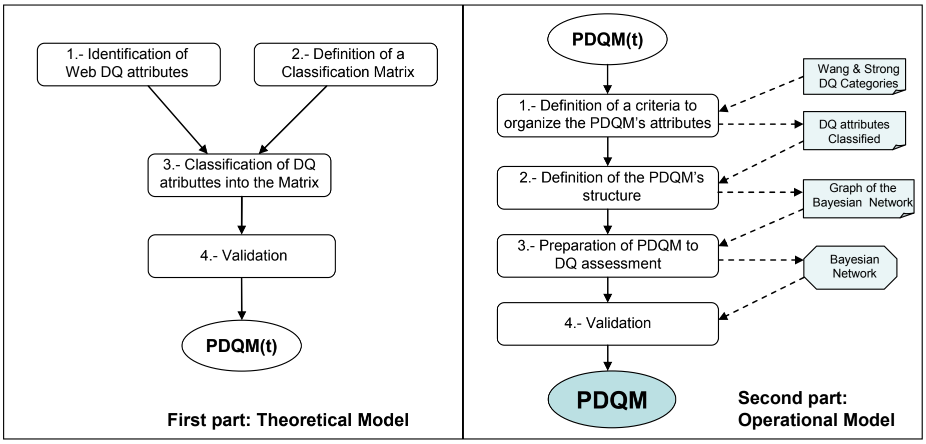


Figure 1. The development process of PDQM.

4 The theoretical definition of PDQM

PDQM is a data quality model for Web portals which focuses on the data consumer's perspective. For the theoretical definition we have considered three key aspects:

- *Data consumer's perspective*. This is represented in our model by the DQ expectations of the data consumer on the Internet, proposed in [27]. These expectations are organized into six categories: Privacy, Content, Quality of values, Presentation, Improvement, and Commitment.
- *Web data quality attributes*. Obtained from DQ frameworks proposed in the literature for different domains in the Web context.
- *Web portal functionalities*. In our model, we used the functionalities that Collins proposes in [8] . These functions are as follows: Data Points and Integration, Taxonomy, Search Capabilities, Help Features, Content Management, Process and Action, Collaboration and Communication, Personalization, Presentation, Administration, and Security.

Based on these aspects, we have developed the model following the four-phase process shown in Figure 1. The theoretical model PDQM(t), made up of a set of 33 DQ attributes; see Table 1, was the end result of this procedure. A more detailed description of the generation of PDQM(t) can be found in [7].

Attractiveness	Documentation	Customer Support
Accessibility	Duplicates	Reliability
Accuracy	Ease of Operation	Reputation
Amount of Data	Expiration	Response Time
Applicability	Flexibility	Security
Availability	Interactivity	Specialization
Believability	Interpretability	Timeliness
Completeness	Novelty	Traceability
Concise Representation	Objectivity	Understandability
Consistent Representation	Organization	Validity
Currency	Relevancy	Value added

Table 1. Data Quality Attributes into PDQM(t).

5 The operationalization of PDQM

If we aim to use PDQM in DQ evaluation processes, we need to convert the theoretical model into an operational model. To do this, we have decided to use a probabilistic approach such as that proposed in [20]. This approach involves Bayesian networks and Fuzzy logic. Bayesian networks (BN) model the problems that involve uncertainty, and combine the advantages of an intuitive visual representation with a sound mathematical basis on Bayesian probability [26]. Fuzzy logic provides an effective conceptual framework for dealing with the problem of knowledge representation in an environment of uncertainty and imprecision [35].

A BN is a directed acyclic graph, whose nodes are the uncertain variables and whose edges are the causal or influential links between variables. A conditional probability function models the uncertain relationship between each node and its parents [26]. In our context, BNs offer an interesting framework with which it is possible to:

- Represent the interrelations among DQ attributes in an intuitive and explicit way by connecting influencing factors to influenced ones.
- Get round the problems of the uncertainty that comes from subjectivity.
- Actually use the network obtained to predict/estimate the DQ of a portal.
- Isolate factors that are responsible for low DQ when that occurs.

Another interesting property of the Bayesian approach is the fact that it considers probability as a dynamic entity that can be updated as more data arrive (self learning mechanism). New data may naturally improve the degree of belief in certain propositions [2]. Consequently, a BN model is particularly adapted to the changing domain of portals.

To convert PDQM into an operational model, we have developed the second part of the process, shown in Figure 1.

5.1 Definition of a criterion to organize PDQM's attributes

As explained in [26], a BN can be built starting from semantically meaningful units called network fragments. A fragment is a set of related random variables that could be constructed and reasoned about separately from other fragments. Thus, a first phase to build a BN for PDQM was to define a criterion that allows us to organize the DQ attributes into a hierarchical structure, with the possibility of creating network fragments.

We have used the conceptual DQ framework developed in [30] as a criterion for organizing the DQ attributes of PDQM In our work, though, we have renamed and redefined the Accessibility category , calling it Operational category. The idea was to emphasize the importance of the role of systems, not only with respect to accessibility and security, but also to aspects such as personalization, collaboration, etc. Having done all this, and taking the definition of each DQ category into account, we have classified all DQ attributes of PDQM into the categories seen below in Table 2. As a result of this, we have defined four network fragments for PDQM, one per category.

DQ Category	DQ Attributes
Intrinsic: It denotes that data have quality in their own right.	Accuracy, Objectivity, Believability, Reputation, Currency, Duplicates, Expiration, Traceability
Operational: It emphasizes the importance of the role of systems; that is, the system must be accessible but secure to allow the personalization and collaboration among other aspects.	Accessibility, Security, Interactivity, Availability, Customer support, Ease of operation, Response time
Contextual: It highlights the requirement which states that DQ must be considered in the context of the task in hand.	Applicability, Completeness, Flexibility, Novelty, Reliability, Relevancy, Specialization, Timeliness, Validity, Value-Added
Representational: It denotes that the system must present data in such a way as to be interpretable and easy to understand, as well as concisely and consistently represented.	Interpretability, Understandability, Concise Representation, Consistent Representation, Amount of Data, Attractiveness, Documentation, Organization

Table 2. Organization of DQ Attributes of PDQM into four DQ Categories.

5.2 Definition of a structure for PDQM

To generate new levels in the BN, we have established relationships of direct influences between the attributes in each category. Our aim was to determine which DQ attribute in a category has direct influence on other attribute(s) in the same category, and eventually on attributes in other category. Each relationship is supported by a premise that represents the direct influence between an attribute and its parent attribute. As an example of how this works, Table 3 shows the relationships established in the DQ Representational.

	Relation of Direct Influence		Premise that supports the direct influence relationships
	Level 2	**Level 3**	
DQ Representational (Level 1)	Concise Representation	-	If data are compactly represented without superfluous elements then they will be better represented.
	Consistent Representation	-	If data are always presented in the same format, are compatible with previous data and consistent with other sources, then they will be better represented.
	Understandability	Interpretability	If data are appropriately presented in language and units for users' capability then they will be understood better.
		Amount of data	If the quantity or volume of data delivered by the portal is appropriate then they will be understood better.
		Documentation	If data have useful documents with meta information then they will be understood better.
		Organization	If data are organized with a consistent combination of visual settings then they will be understood better.
	Attractiveness	Organization	If data are organized with a consistent combination of visual settings then they will be more attractive to data consumers.

Table 3. Relationships in the Representational DQ Category.

Taking these relationships (in the four categories) as our basis, we have built the graph of the BN which represents PDQM, see Figure 2.

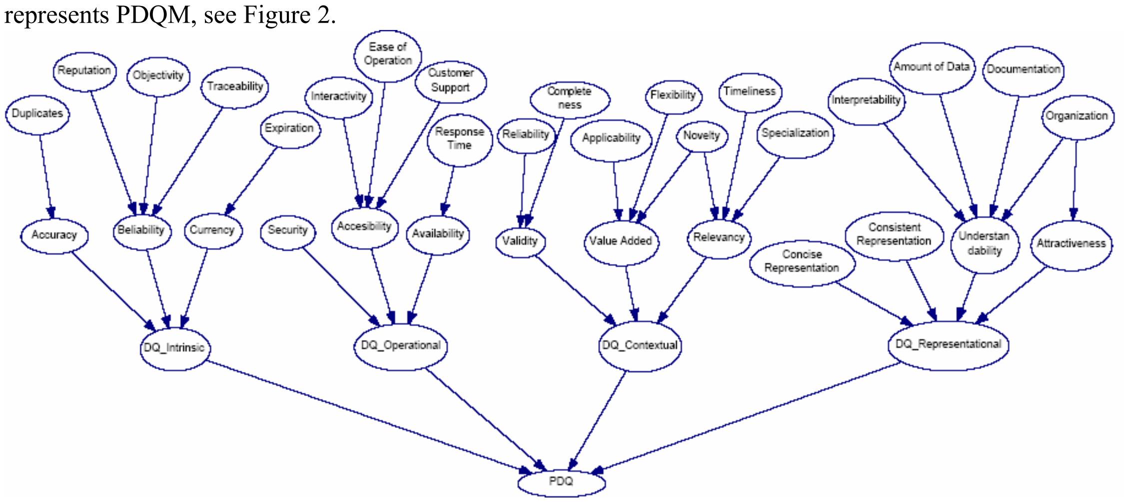


Figure 2. Graph of the BN that represents the PDQM.

5.3 Preparation of DQ_Representational fragment for DQ assessment

Although our final objective is to create a comprehensive BN model for PDQM, taking advantage of the possibility of working separately with each fragment in the BN, we have decided to start working with the DQ_Representational fragment. Thus, in the remainder of this paper we will tackle only the definition of the DQ_Representational fragment. However the activities developed to prepare this fragment will be the same as those that will be used to prepare the rest of the fragment in the PDQM. To prepare the sub-network for assessing the DQ, the following sub-phases were developed:

1) The creation of synthetic nodes to simplify the fragment if necessary, i.e., to reduce the number of parents for each node.
2) The definition of quantifiable variables for the entry nodes (input nodes) in the fragment.
3) The definition of the Node Probability Table (NPT) for each node in the fragment.

Another consideration when developing this phase, applicable to sub-phase 3, was that the BN had to be prepared to evaluate Web portals in specific domains, because data consumers can have different demands and needs, depending on the particular portal domain [29]. In this case, we have selected the domain of university Web portals.

5.3.1. Creation of synthetic nodes.

The original fragment had two nodes with four parents (Understandability and DQ_Representational) and two synthetic nodes (Representation and Volume of Data) had to be created in order to reduce the combinatory explosion in subsequent steps in the preparation of the NPTs. See in Figure 3 the graphs 1 and 2.

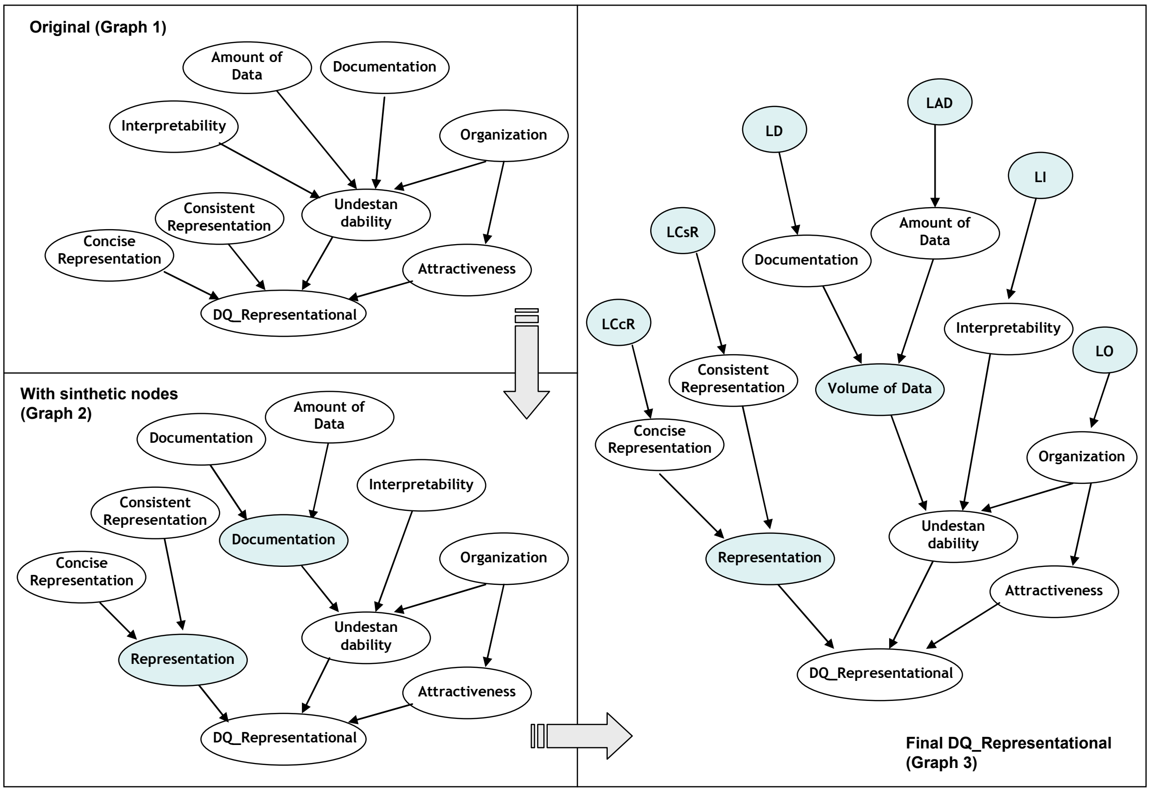


Figure 3. Sub-network DQ_Representational.

5.3.2. Definition of quantifiable variables.

In this sub-phase measures for the quantifiable variables (input nodes) in the sub-network were defined. To define and describe the measures in a more formal way, we have used the SMO (Software Measurement Ontology) proposed in [3]. This ontology provides a set of common measurement concepts, among them being: measure, scale, type of scale, unit of measurement, base measure, derived measure and indicator.

Thus we have defined an indicator for each input node in the fragment (see Figure 3, leaf nodes of the graph 3). To calculate each indicator, we have defined several measures that will be automated and calculated for a given portal. The indicators defined will take a numerical value from between 0 to 1. As the number of possible values for each input node can be infinite, we have transformed them into discrete variables. This is done to ease the definition of probabilities. According to [20], this transformation can be achieved using fuzzy logic. Hence, we have defined for each indicator a membership function that transforms the value of the indicator into a set of probabilities, each of them corresponding to a label/class.

To show how this process works, we next explain the definition of the LCsR (Level of Consistent Representation) indicator, which generates a measure for the input node Concise Representation. The measures selected for this attribute take as their focal point the consistency of the format and the compatibility between the pages in the portal. This is not only because these aspects are more obvious for data consumers when this attribute is evaluated, but also because they can be measured in an objective way. We have defined measures based on the use of Style on the pages of the Web portal for this indicator, as well as on the correspondence between a source page and the destination pages.

One type of correspondence measured, for instance, was if the text associated with a link was repeated on a destination page. We have defined a set of base and derived measures to carry out that task. Taking these measures as a starting point, we gone on to define their Analysis Model [3], which includes a Formula that gives us a numerical value and a Decision Criteria in the form of a membership function; see Table 4.

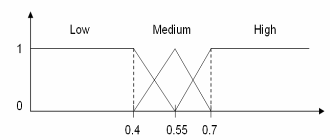
LCsR (Level of Consistent Representation)	
Formula	**Decision Criteria**
LCsR = PSSD*0.5 + SDCD*0.5	
Derived Measures	
PSSD: Pages with the same Style SDCD: Source and Destination Correspondence	

Table 4. Analysis model for LCsR indicator.

5.3.3. Definition of node probability tables

For each node in the BN, its probability distribution was defined. The node probability table for the input nodes is derived directly from the value taken by the corresponding indicator. The intermediate nodes are ones which are defined by their parents and they are not directly measurable. The definition of node probability tables in our work was given on the basis of judgment afforded by experts, taking into account the Web portal domain selected (University portals) as well. Following on with our example, in Table 5 we can see the NPT for the Consistent Representation of the node.

LCsR	Low	Medium	High
Bad	0.9	0.05	0.01
Medium	0.09	0.9	0.09
Good	0.01	0.05	0.9

Table 5. Node probability table for Consistent Representation

The process, then, is as follows. For a given University Web portal we calculate the PSSD and SDCD measures associated with the indicator LCsR. When we apply the formula defined, we will obtain a value between 0 and 1. Using the Decision Criteria, we will transform this value into a set of probabilities for the labels “Low”, “Medium” and “High”. With this value, and using the probability table shown in Table 5, we can derive the value of the Consistent Representation- for example, the probability of having a consistent representation “Bad” when LCsR takes a “Low” value is the 0,9). So this is when we will have “evidence” that will be propagated, via a causal link, to the child nodes of Consistent Representation- in this case the node Representation (see Figure 3 graph 3).

This process is similar for the rest of the input nodes in the sub-network. In general, the idea is that values for input nodes for a given Web portal are measured directly. These values are transformed into a set of probabilities, each corresponding to a label/class. These probability values which are calculated for each input node are known as “evidence”. This evidence propagates through the BN via causal links until the level of representational DQ in the Web portal is obtained.

5.3.4. Validation of PDQM

Based on the measures defined for the fragment DQ Representational, we have development of a tool with which to calculate this. The tool asks for the URL of a portal, downloads it and applies the defined measures. The results are transformed and entered in the input nodes; after this, the BN generates the DQ evaluation.

To validate PDQM, we conducted an experiment with which to obtain the judgments of a group of data consumers about the DQ in a university Web portal. The idea was to compare their valuations with the assessment made with PDQM for the same portal. As a result we have obtained the subject and PDQM scores shown in Table 6.

	Low	Medium	High
Subjects	0.17	0.68	0.16
PDQM Tool	0.18	0.58	0.24

Table 6. Results obtained in the Validation phase

As a conclusion of this validation, we can say, in general, that the results of PDQM are close to the subject´s valuations. This is because with both evaluation methods we have obtained a Medium level for the DQ Representational.

However, we can observe a significant difference between the scores for the medium and high categories. We believe that the main reason for this is the extreme values obtained for some indicators and thus, we must review their definitions and the form by which they are calculated. But this is not the only reason. Together with the former problem, we believe that the design of the Web portal evaluated may also influence this result. For example, in order to be able to calculate the Level of Amount of Data (LAD indicator LAD in figure 3, graph 3), it is necessary to know the distribution of words per page. The measured portal presents values for this measure which can be considered as outliers (they take extreme values that do not follow a uniform distribution). Obviously, these values need to be removed from the calculation of the measure. Because of this we are now refining the calculations made by the tool by detecting and eliminating the outliers in our measures.

Of course we also need to repeat the validation process by including a great number of Web portals, in order to be sure that the BN accurately estimates the Representational DQ of any portal.

6 Conclusions

In this paper we have presented the work we have done on the development of a DQ model for Web portals. In our previous work we had defined a theoretical model. This is formed from a set of 33 DQ attributes that can be used for DQ evaluation in Web portals. The selection of these attributes was guided by the data consumer’s perspective. In the second part, we have begun the transformation of the theoretical model into an operational one, using a probabilistic approach.

The choice of a probabilistic approach to generate the framework is motivated by the fact that by using it we get around many problematic issues in quality assessment: threshold value definition, measure combination and uncertainty.

This paper has described the process of preparation of a fragment of PDQM for its assessment in a specific domain of a Web portal. Measures and distribution of probabilities have been defined.

The most important contribution of this work is the organization of the PDQM attributes into a BN that represents a generic structure which can be adapted to both the goal and the context of evaluation.

Starting from the perspective of the goal, users can choose the fragment or sub-network that evaluates the characteristics they are interested in. From the point of view of the context, the parameters (probabilities) can be changed, in such a way as to consider the specific context of the evaluated portal.

As future work, we first plan to repeat the validation process in order to be sure that the BN accurately estimates the Representational DQ of any portal. After this we will work to complete the operationalization of the other fragments of PDQM.

Acknowledgements

This research is part of the following projects: ESFINGE (TIC2006-15175-C05-05) granted by the Dirección General de Investigación del Ministerio de Ciencia y Tecnología (Spain), CALIPSO (TIN20005-24055-E) supported by the Ministerio de Educación y Ciencia (Spain), DIMENSIONS (PBC-05-012-1) supported by FEDER and by the "Consejería de Educación y Ciencia, Junta de Comunidades de Castilla-La Mancha" (Spain) and COMPETISOFT (506PI0287) financed by CYTED.

References

[1] M. Aboelmeged. A Soft System Perspective on Information Quality in Electronic Commerce. in Proceeding of the Fifth Conference on Information Quality. 2000.

[2] P. Baldi, P. Frasconi, and P. Smyth, Modeling the Internet and the Web; Probabilistic Methods and Algorithms. 2003: Wiley.

[3] M. Bertoa, F. García, and A. Vallecillo, An Ontology for Software Measument, in Ontologies for Software Engineering and Software Technology, C. Calero, F. Ruiz, and M. Piattini, Editors. 2006.

[4] M. Bouzeghoub and Z. Kedad, Quality in Data Warehousing, in Information and Database Quality, M. Piattini, C. Calero, and M. Genero, Editors. 2001, Kluwer Academic Publishers.

[5] M. Burgess, N. Fiddian, and W. Gray. Quality Measures and The Information Consumer. in Proceeding of the Ninth International Conference on Information Quality. 2004.

[6] C. Cappiello, C. Francalanci, and B. Pernici. Data quality assessment from the user´s perspective. in International Workshop on Information Quality in Information Systems, (IQIS2004). 2004. Paris, Francia: ACM.

[7] A. Caro, C. Calero, I. Caballero, and M. Piattini, A Theoretical Model doe Portal Data Quality Data Consumer Oriented. Data & Knowledge Engineering. Sent to be published., 2007.

[8] H. Collins, Corporate Portal Definition and Features. 2001: AMACOM.

[9] M. Eppler, Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes. 2003: Springer.

[10] M. Eppler, R. Algesheimer, and M. Dimpfel. Quality Criteria of Content-Driven Websites and Their Influence on Customer Satisfaction and Loyalty: An Empirical Test of an Information Quality Framework. in Proceeding of the Eighth International Conference on Information Quality. 2003.

[11] A. Even, G. Shankaranarayanan, and S. Watts. Enhancing Decision Making with Process Metadata: Theoretical Framework, Research Tool, and Exploratory Examination. in 39th Annual Hawaii International Conference on System Sciences (HICSS'06). 2006.

[12] M. Gertz, T. Ozsu, G. Saake, and K.-U. Sattler, Report on the Dagstuhl Seminar "Data Quality on the Web". SIGMOD Record, 2004. vol. 33, Nº 1: p. 127-132.

[13] G. Graefe. Incredible Information on the Internet: Biased Information Provision and a Lack of Credibility as a Cause of Insufficient Information Quality. in Proceeding of the Eighth International Conference on Information Quality. 2003.

[14] E. Herrera-Viedma, G. Pasi, and A. Lopez-Herrera, Evaluating the Information Quality of Web Sites: A Quality Methodology Based on Fuzzy Computing with Words. Journal of American Society for Information Science and Technology, 2006. 54(4): p. 538-549.

[15] P. Katerattanakul and K. Siau, Information quality in internet commerce desing, in Information and Database Quality, M. Piattini, C. Calero, and M. Genero, Editors. 2001, Kluwer Academic Publishers.

[16] P. Katerattanakul and K. Siau. Measuring Information Quality of Web Sites: Development of an Instrument. in Proceeding of the 20th International Conference on Information System. 1999.

[17] S.A. Knight and J.M. Burn, Developing a Framework for Assessing Information Quality on the World Wide Web. Informing Science Journal, 2005. 8: p. 159-172.

[18] Y. Lee, AIMQ: a methodology for information quality assessment. Information and Management. Elsevier Science, 2002: p. 133-146.

[19] M. Mahdavi, J. Shepherd, and B. Benatallah. A Collaborative Approach for Caching Dynamic Data in Portal Applications. in Proceedings of the fifteenth conference on Australian database. 2004.

[20] G. Malak, H. Sahraoui, L. Badri, and M. Badri. Modeling Web-Based Applications Quality: A Probabilistic Approach. in 7th International Conference on Web Information Systems Engineering. 2006. Wuhan, China: Springer LNCS.

[21] C. Marchetti, M. Mecella, M. Scannapieco, and A. Virgillito. Enabling Data Quality Notification in Cooperative Information Systems through a Web-service based Architecture. in Proceeding of the Fourth International Conference on Web Information Systems Engineering. 2003.

[22] H. Melkas. Analyzing Information Quality in Virtual service Networks with Qualitative Interview Data. in Proceeding of the Ninth International Conference on Information Quality. 2004.

[23] M.Á. Moraga, C. Calero, and M. Piattini, Comparing different quality models for portals. Online Information Review., 2006. Vol. 30(5): p. 555-568.

[24] V. Moustakis, C. Litos, A. Dalivigas, and L. Tsironis. Website Quality Assesment Criteria. in Proceeding of the Ninth International Conference on Information Quality. 2004.

[25] F. Naumann and C. Rolker. Assesment Methods for Information Quality Criteria. in Proceeding of the Fifth International Conference on Information Quality. 2000.

[26] M. Neil, N.E. Fenton, and L. Nielsen, Building large-scale Bayesian Networks. The Knowledge Engineering Review, 2000. 15(3): p. 257-284.

[27] T. Redman, Data Quality: The field guide. 2000, Boston: Digital Press.

[28] C. Schaupp, W. Fan, and F. Belanger. Determining Success for Different Website Goals. in 39th Hawaii International International Conference on Systems Science (HICSS-39 2006). 2006. Kauai, HI, USA: IEEE Computer Society.

[29] G. Shanks and B. Corbitt. Understanding Data Quality: Social and Cultural Aspects. in 10th Australasian Conference on Information Systems. 1999. Wellington, New Zealand.

[30] D. Strong, Y. Lee, and R. Wang, Data Quality in Context. Communications of the ACM, 1997. Vol. 40, Nº 5: p. 103 -110.

[31] R. Wang, A Product Perspective on Total data Quality Management. Communications of the ACM, 1998. Vol. 41, Nº 2: p. 54-65.

[32] R. Wang and D. Strong, Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems; Armonk; Spring 1996, 1996. 12(4): p. 5-33.

[33] L. Xiao and S. Dasgupta, User Satisfaction with Web Portals: An empirical Study, in In Web Systems Design and Online Consumer Behavior, Y. Gao, Editor. 2005, Idea Group Publishing, Hershey, PA. p. 193-205.

[34] Z. Yang, S. Cai, Z. Zhou, and N. Zhou, Development and validation of an instrument to measure user perceived service quality of information presenting Web portals. Information and Management. Elsevier Science, 2004. 42: p. 575-589.

[35] L.A. Zadeh, Knowledge Representation in Fuzzy Logic. IEEE Transactions on Knowledge and Data Engineering, 1989. 1(1): p. 89-100.

[36] Y. Zhu and A. Buchmann. Evaluating and Selecting Web Sources as external Information Resources of a Data Warehouse. in Proceeding of the 3rd International Conference on Web Information Systems Engineering. 2002.

ÍNDICE DE AUTORES

A continuación se presentan ordenados por nombre de autor los números de página de los artículos de los eventos de CLEI2007. A la derecha se anota el número de página.

A

D

E

F

G

H

M

N

S

T

Auspiciadores ORO

RACSA
Líderes en Internet y Conectividad Empresarial

Auspiciadores PLATA

BANCO NACIONAL
Microsoft
CRUSA X ANIVERSARIO

Auspiciadores BRONCE

BCR
Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura
intel Leap ahead
PC WORLD
cisco
UNIVERSIDAD INTERAMERICANA
Laureate International Universities
GBM An IBM alliance company
ITNOW

Auspiciadores Jade

BOLSA NACIONAL DE VALORES
Copa Airlines copaair.com
BAC San
CREDOMATIC Siempre es posible
AVANTICA TECHNOLOGIES Creating Solutions
AURA INTERACTIVA
PanGenesis Software Development, Maintenance, Translation
P&G
L.L.Bean
EF EL FINANCIERO