Data integration: Quality aspects

It is now possible to observe that the advances in technology have led to increased capacity to generate and store huge amounts of data in all areas of knowledge, characterizing a generalized explosion of data. In the area of electric power, the vast amount of data collected by different systems of supervision and control and stored in historical bases, has become a potential source of implicit knowledge of great value to the management of electrical systems. The availability of huge volumes of data, from different sources with different levels of quality, makes the extraction of this knowledge time consuming, costly and often ineffective. Additionally, there is the need to evaluate the quality of primary data between those historians, since they may present incomplete or inconsistent due factors related with industrial processes of acquisition: Measuring instruments with bias, uncalibrated or malfunctioning: Manual entries of incomplete data, illegible or not made: Stops and starts of equipments: Abnormal behavior of the process In this paper we consider the search for quality in the user's perspective of the information ("fitness for use" principle) where the definition of the level of an adequate quality of data depends on the context. We seek to verify if information currently existing in systems providing data are sufficient to ensure the quality of their data. This check is made by mapping the information in metrics of the quality of data. For this study, an adaptation of a methodology was done with the following steps: identification of the system being evaluated; identification and selection of the most important data sets; evaluation of the quality of selected system and finally, the quality measurement of their data. The first step was characterized by the choice among the various systems, the system being the object of evaluation, considered in this study as the Events Repository System. This system consists of a set of computational modules to support t- > - > he tasks of analysis and monitoring of large industrial processes. The quality step of the information system focused verifying the functionality of this system in the aspects of data storage and retrieval. Several dimensions of quality relevant to the intended application of data, were defined and assessed in this step. With these dimensions defined it was possible to check the quality of the stored data, and point out the actions and precautions to be taken to ensure the quality that is required.

This paper appears in: Transmission and Distribution Conference and Exposition: Latin America (T&D-LA), 2010 IEEE/PES, Issue Date: 8-10 Nov. 2010, Written by: Bastos, M.R.; Martini, J.S.C.; de Almeida, J.R.; Viana, S.

SECTION I
Introduction

The quality of the information currently considered as a product in the production system [1], is essential for the survival of the organizations; but should be defined, measured, analyzed and enhanced continuously to meet the needs of their consumers [2].

In the electricity area, the immense amount of data collected by different supervision and control systems and stored later in historical bases, is a potential source of implicit knowledge of great value for the management of electrical systems.

The availability of these huge volumes of data from different sources, with different levels of quality, makes the extraction of this knowledge become time consuming, costly and often ineffective.

The measurement of quality, of this massive amount of information becomes a way of classifying and identifying information that may better suit the interests and needs of users [1].

Considering information as the result of an organization, processing and/or data analysis, the deficiencies that affect their quality, can originate at any point in this process, including even the quality of their data.

once the decisions are made based on a given set of information, available during the process of decision-making, the negative impacts due to problems with information quality are translated into unnecessary costs, with decisions whose application may not produce the desired results.

It is expected that, although involving some degree of uncertainty, decisions based on more relevant, complete, accurate and updated data, have better chances of better results [3].

This paper presents a study of quality of data to be used by a System of “Data Warehouse Fuzzy” (SDWF). The data used by the SDWF come from three other computer systems and interviews/questionnaires with the operations teams (specialist information). Errors can occur in any of these systems due to data entry (typing, input latency from the user, field devices communication faults, noise in measurements, inconsistent input values, etc.)

The intent here is to verify the information currently existing in the systems providing data are sufficient to ensure the quality of their data. This check is done by mapping these information in metrics of data quality in a user's perspective of the information (principle of “fitness for use”), where the definition of an adequate quality of data depends on the context of its use [4] [5].

Since the value of a system is determined by the quality of the information it processes [1], the need to check the quality of data being used by SDWF is justified. The quality of information is a prerequisite for the deployment of information systems, because without this attribute, resulting information of these systems will also present low quality.

In addition, considering that data provides information that takes to knowledge [6], and that the new information can be generated from the existing information; the here adopted approach can be generalized both for data and information

SECTION II
QUALITY

Quality “per se” is a concept of major importance in the organizational field; having as one of its definition: the measure of fitness for purpose. Thus, to a product or service have quality, their characteristics must meet the expectations of their customers.

Considering information as a product that should meet the expectations of its user, the concept of quality can be extended also to quantify the quality of the data from which information is generated.

User especifications, when defining their needs and context of use, provide greater ease of measuring data quality, since establishes a standard, a goal for which is possible to measure the degree of compliance of the data [1].

Additionally, in this same point of view, the degree of adherence between the views presented by an information system and the same data in a real world, can also be used to quantify the quality of data [8].

A Dimensions of Data Quality

The data, as raw material for information systems, unlike the raw material in physical processes, are not exhausted, they can be reused repeatedly in order to meet different goals [8].

The possibility of this use, of same data, by different users on different objectives, makes the quality of this data be presented in distinct ways (different perceptions of the same concept) for their various users, meeting or not their requirements and individual expectations [9].

This multidimensional nature of the concept of quality, together with the requirement that data must comply effectively and efficiently, to a number of dimensions that ensure their quality in attendance to a specific intended use [7], requires the individual assessment of these dimensions to determine the quality.

The definition of these dimensions or their aspects related to the requirements of the specific use of the data, characterized as a difficult task, which once done, will allow the direct definition of metrics for quality assessment [5].

SECTION III
“DATA WAREHOUSE FUZZY”(SDWF) SYSTEM

The SDWF system, subject of this study, is based on a platform composed by sources of information (data origins used by SDWF), by modules of treatment of these data, module of knowledge modeling and modules of data and rules repositories, according to Fig.1.

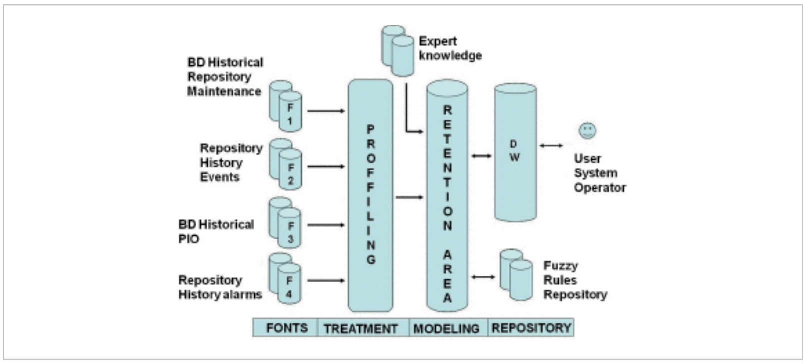

Fig. 1. Platform of "Data Warehouse Fuzzy" System (SDWF)

A. Data Provider Systems for SDWF

The SDWF uses information from multiple systems, whose data structures are completely different; with formatting rules, data fields, data entry mode and data type with their own particularities and conventions. By SDWF there is no quality control (primary data) of data supplied by these systems.

1) Maintenance Management System (F1, in Fig.1)
This system is directly used by maintenance crews, for information records on equipments in field operations. The identification information of the equipment in maintenance, maintenance team members, report of problem and solution, maintenance period, time of service travel, travel costs and others, are inserted in this system as a subsidy for maintenance management.

2) Events Repository System (F2, in Fig.1)
This system is composed by the events historian, which stores the occurrences of the events related to the various devices constituting the electrical system. Stored information include the states of equipments (on, off, locked, unlocked, etc.) and measurements of analog quantities (current, voltage, power, temperature, frequency, etc.). These information are used for post-operation analysis, reports preparation and studies of trends among others.

All information are marking the instant of its occurrence.

3) Operation Management System (F3, in Fig.1)
All management of the operation of the Transmission System is made through this computational tool. All actions relating to the release of equipments for interventions, equipment energizing release, etc, are entered in it. The information of equipment identification, intervention period, type of intervention, reason for intervention, those responsible for request and services and other information are inserted by programming, support and maintenance teams.

4) Alarms Repository System (F4, in Fig.1)
The alarms repository concentrates the information related to the occurrences of the equipments state variations, operational limits exceeded, equipments operational conditions (in fault, normal, etc.), computational system alarms, communication links losses, etc.

All information are marking the instant of its occurrence.

5) "Fuzzy" Rules Repository
This repository contains the several "fuzzy" rules that are apply to a set of information, so producing the final information delivered to the DWF user.

Sdwf Functional Modules

Functionally the SDWF consists of Profiling, Retention Area and DW modules.

1) Profiling Module
The process of data treatment characterized by this module is defined as the application of analysis techniques with regard to knowledge of the content, structure and quality of current data (7).

In this module is made the detection of the anomalies that data can provide in a database. In this process of "data profiling" one seeks to systematically detect the errors, inconsistencies, redundancies and the existence of incomplete information in data and respective metadata.

2) Retention Area Module
The process of modelling involves the processing operations and source data cleaning (F1 to F4), involving correction of syntactic errors and data preparation for integration. This module is a data storage area, where they undergo the necessary changes, and then be loaded in the "Data Warehouse" Module. These changes may include many actions, such as: data cleaning, data deletion and combination of the various data sources. In this module, the integration of the representation of the specialist knowledge to other data is also treated, generating information to be stored by the repository, DW Module.

3) DW Module
Is the ultimate source of information to the SDWF user, which will provide the resulting information product of the "Fuzzy" correlation and inferences from various information, considered together.

C. Metering Data Quality

The absence from the SDWF, of quality control of the data from multiple source systems, together with the fact that the quality of its information product depends on the quality of all of its inputs (data), indicates a need for quality assessment of its primary data.

This task of checking for possible problems in the data is compromised by the lack of explicit metadata in primary data structures, leaving as an alternative to analyze if existing information on each of these systems can be used to assess the quality of their own data.

For this study, the methodology presented by [1] is adapted, with the following steps: identification of the System to be assessed, identification and selection of most important Data sets, assessment of the Quality of this selected Information System (QSI) and finally the metering of Quality of its Data (QD), Fig.2.

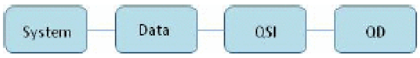

Fig. 2. Steps of the process of data quality metering. Adapted from [1].

1) System Selection - Events Repository System
The first step was characterized by the choice among the various systems, of the system being the object of evaluation, considered the Events Repository System in this study.

This system consists of a set of computational modules to support the analysis and monitoring of large industrial processes.

A key point of this system is its ability to compress and store data efficiently, so this efficiency depends on its proper parameterization.

It performs the tasks of collecting, storing and recovering data in numeric or string format, also acting as a Server for client applications based on MS Windows@.

The process data are collected and stored together with their time labels (time at which the event occurred). For each process variable to be a historian, a historical point is defined, containing approximately 50 attributes. These attributes were kept in the database of the points, are used to define how the data will be collected and archived, directly impacting on the efficiency of handling this data by the historian.

Since the definition of the attributes of a point (data) was made so that there is no loss of information in storage, the other attributes can be used for optimization of the historical process.

2) Information System Quality (QSI)
The quality dimensions adopted for the events repository are related in Fig.3, which are sufficient to meet the requirements of the users of these data.

Considering the focus on the quality of data is something perceived by the users and that the adherence to its use of data is a measure of this quality, the crucial dimension of quality in this work was the **Relevance** of the data, given by the user.

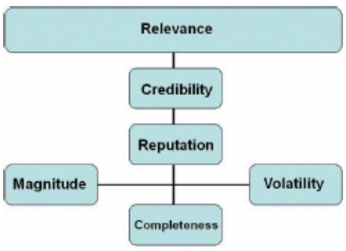

Fig. 3. Data dimensions for the Events Repository

Dimensions adopted and their meanings are:

Relevance	how much the data is applicable and useful for the task
Credibility	how much the data is accepted as true and authentic
Reputation	credibility of data source
Completeness	all data necessary are present
Size	range of allowed values – minimum value and maximum value
Volatility	frequency of allowed change

Since all analyzed data were those selected by the user, all have maximum degree of relevance to the proposed study. The relevance could also be evaluated by using the “simple ratio” operator (5) given to each of data present in the repository to be useful for the task of examining the state proposal.

From the characterization of the relevance of the data, it has its quality mapped in the dimensions of credibility, reputation, size, completeness and volatility.

a) Data Structure of the Events Repository

The database of the events repository (historical events) can be characterized by the structure presented in Fig.4, which references only the attributes considered here.

Where the attributes have the following characterization:

Fig. 4. Partial structure of the database of the events repository.

time:	time stamp, in which the event has occurred
tag:	unique identification of the event (primary key)
value:	contains the record of the size presented (measurement) by the event
deviation:	parameter that defines the degree of the variation of the value, from which the event is recorded in the database
excmin:	difference in the minimum time that the event is not recorded, even exceeding the deviation
excmax:	difference in the maximum time that the event is recorded (even not exceeding deviation)
zero:	minimum value to measure the event
span:	value of the total allowed range to measure the event
quality :	indicative of the event quality (related to its acquisition) and of its source
type:	type of the event (round, floating point, string)

b) Mapping of the Information in Quality Dimensions

The initial hypothesis is that data from historical events have several additional information that can be used to verify the consistency of these stored data, following rules for data quality.

Analyzing each of the attributes of the structure shown in Fig.4, their mapping could be made in relation to the dimensions of quality previously related, this mapping being represented in Fig.5.

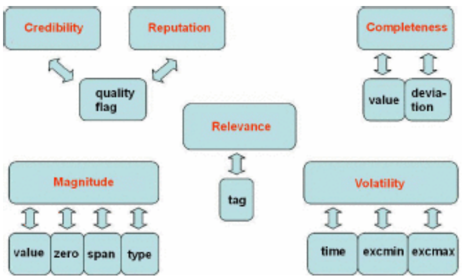

Fig. 5. Mapping of data attributes of the events repository

The **Credibility** and **Reputation** dimensions are related to the primary source (origin) of the data (field measurements and operator manual input) [1]. These dimensions are mapped by the **quality flag** attribute, indicating if the data is valid (from its source), also informing if it comes from automatic acquisition or manual input. Automatic acquisition was considered, in this work, with a better reputation than the manual input.

The **Completeness** dimension is mapped by the **value** and **deviation** attributes (equivalent to the concept of dead band), so if the deviation is properly parameterized, no loss of information (value) will occur in the data storage and retrieving process.

Regarding the **Magnitude** scale, this is mapped by the attribute **value** of data, in engineering units (MVA: apparent electric power in millions of Volt-Ampere and A: electric current in Amperes) characterized by the attributes **zero, span** and **type.** The **value** must be within the range defined by the attributes **zero** and **span,** besides having the right type of magnitude, since only the analog type (real, floating point) has the **zero** and **span** attributes.

The resulting mapping for the **Volatility** scale is characterized by the attributes: **time, excmin** and **excmax.** The variation rate of the values of considered data must be within the limits indicated by the attributes **excmin** and **excmax.**

For each dimension, the composition of the attributes used set whether the attribute **value** is relevant and has allowed values, within the limits of defined periodicity and variations, as if the recorded data in the attribute value is consistent with its type and presents the appropriate **quality flag**.

Among the tens of thousands of historian points, were selected the related to the load (measured in MVA or using de measurement of the A current) in power transformers. This choice is justified by the importance of this information to determine the load distribution through the Electric Transmission System.

In Fig.6 one can see a partial relation of the attributes of multiple transformers of different classes of voltage and power.

tag	units	zero	span	pointtype	excdev	excmin	excmax
AGV .13.8 TR-7 AMPS .AV	AMPS	-8	531,9000244	Float32	0,188005256	0	600
AGV .440 TR-4 MVA C .AV	MVA	-750	1500	Float32	0,066666667	0	600
AGV .440 TR-5 MVA C .AV	MVA	-305	610	Float32	0,163934426	0	600
AGV .AGV 440TR-8 MVA .AV	MVA	-750	1500,5	Float32	0,066644452	0	600
AGV .AGV TR-8 AMPS .AV	AMPS	0	1500	Float32	0,066666667	0	600
ANH .230 TR-4 FASE B A .AV	AMPS	0	1255	Float32	0,079681275	0	600
ANH .230 TR-4 MVA .AV	MVA	0	500	Float32	0,2	0	600
APAT .88 TR-1 A .AV	AMPS	0	394	Float32	0,253807107	0	600
APAT .88 TR-1 MVA .AV	MVA	0	60	Float32	1,666666865	0	600
APAT .88 TR-2 A .AV	AMPS	0	1000	Float32	0,1	0	600
APAT .88 TR-2 MVA .AV	MVA	0	60	Float32	1,666666865	0	600
APAT .88 TR-3 A .AV	AMPS	0	394	Float32	0,253807107	0	600
APAT .88 TR-3 MVA .AV	MVA	0	60	Float32	1,666666865	0	600
ARA .138 TR-1 FASE A A .AV	AMPS	1	1599	Float32	0,062539087	0	600
ARA .138 TR-1 MVA .AV	MVA	-1147	2294	Float32	0,043591979	0	600
ARA .138 TR-2 FASE A A .AV	AMPS	1	1599	Float32	0,062539087	0	600
ARA .138 TR-2 MVA .AV	MVA	-1147	2294	Float32	0,043591979	0	600
ARA .138 TR-3 MVA .AV	MVA	-179	358	Float32	0,279329609	0	600
ARA .138 TR-4 FASE B A .AV	AMPS	1	1999	Float32	0,050025013	0	600
ARA .138 TR-4 MVA .AV	MVA	-1434	2868	Float32	0,034867503	0	600
ARA .440 TR-1 MVA .AV	MVA	0	783	Float32	0,127713921	0	600
ARA .440 TR-2 FASE A C .AV	AMPS	0	2000	Float32	0,05	0	600
ARA .440 TR-2 MVA .AV	MVA	-799	1582	Float32	0,063211125	0	600
ARA .440 TR-4 FASE A B .AV	AMPS	0	2000	Float32	0,05	0	600
ARA .440 TR-4 MVA .AV	MVA	-799	1582	Float32	0,063211125	0	600
ASS .230 TR-2 MVA .AV	MVA	0	1524	Float32	0,065616798	0	600
ASS .230 TR-3 MVA .AV	MVA	-1	39	Float32	2,564102564	0	600
ASS .230 TR-4 MVA .AV	MVA	-38	76	Float32	1,315789474	0	600
ASS .440 TR-2 MVA .AV	MVA	-336	672	Float32	0,148809524	0	600
ASS .440 TR-5 MVA .AV	MVA	-1432	2864	Float32	0,034916201	0	600
ASS .525 TR-5 MVA .AV	MVA	-1432	2864	Float32	0,034916201	0	600
ASS .ASS 230TR-4 MVA .AV	MVA	-0,5	39,5	Float32	2,53164557	0	600

Fig. 6. Partial relation of the attributes of the power transformers

It may be noted the some of the attributes values are “default”, which signals a possible optimization of the historical process, through a more detailed assessment of the adopted parameterization. For example, the attribute **excmax** which restricts the maximum rate of change of the data. It determines how fast the data can be changed and still be considered valid for storage.

By defining the value for this attribute is considered the physical nature of the data, its possible behavior, hysteresis, etc. Thus, data with variation above this limit will not correspond to the reality of the process, characterizing a measurement/acquisition noise.

As for the assessment of data quality, the consistency of the multiple attributes and the consistency of the measured quantity, for example: measurement within the limits from **zero** and **span**, Fig.7. and Fig.8.

	AGV .440 TR-4 MVA C .AV
data points (maximum reached):	64000
06-fev-07 13:12:30	Pt Created
08-fev-07 01:15:26	430,5589294
08-fev-07 01:16:34	431,7683411
08-fev-07 01:19:44	425,4699402
08-fev-07 01:21:02	420,7043152
08-fev-07 01:24:58	430,7239075
08-fev-07 01:26:08	425,5863647
08-fev-07 01:26:50	430,7239075
08-fev-07 01:27:38	430,7239075
08-fev-07 01:33:12	430,453949
08-fev-07 01:34:28	425,1966553
08-fev-07 01:34:50	431,7683411
08-fev-07 01:36:28	432,8393555
08-fev-07 01:38:42	439,4154053
08-fev-07 01:39:56	442,6152649
08-fev-07 01:41:04	447,3072815
08-fev-07 01:42:12	439,326416
08-fev-07 01:50:46	442,6152649
08-fev-07 01:54:44	438,0108948
08-fev-07 01:55:38	427,0663452
08-fev-07 01:58:38	434,9544983
08-fev-07 02:00:36	425,1966553
08-fev-07 02:00:52	427,0663452
08-fev-07 02:02:58	425,3029175
08-fev-07 02:03:38	409,7111511
08-fev-07 02:05:44	409,8320313
08-fev-07 02:06:12	409,8320313
08-fev-07 02:06:42	411,1446838
08-fev-07 02:07:32	395,7275085
08-fev-07 02:08:24	409,8320313
08-fev-07 02:09:22	398,8703613
08-fev-07 02:10:00	406,7412415
08-fev-07 02:10:56	411,1446838

Fig. 7. Values of Power MVA TR-4 440kV AGV

Tag	units	zero	Span	pointtype
AGV .13.8 TR-7 AMPS .AV	AMPS	-8	531,9000244	Float32
AGV .440 TR-4 MVAC .AV	MVA	-750	1500	Float32
AGV .440 TR-5 MVAC .AV	MVA	-305	610	Float32
AGV .AGV 440TR-8 MVA .AV	MVA	-750	1500,5	Float32

Fig. 8. Attibutes

SECTION IV.
CONCLUSION

The study concludes that the Events Repository already presents in its original data structure, additional information that enables a multidimensional assessment of the quality of the value attribute, regarded as the primary data to be provided, by this module, for the use of the SDWF computational processes.

The appropriate parameterization of the various attributes requires the knowledge of the nature and behavior of the **value** attribute, thus constituting a model of it. By incorporating these information, these attributes may then be considered as metadata of the **value** attribute.

The mapping of these metadata on the dimensions studied here has allowed the assessment of the quality, within the context of the attendance of data to its intended use. Within this focus, in choosing the dimensions of the quality, the data user was considered as having the basic understanding of information quality, needed to perform the work.

The quality of the **value** attribute could be evaluated through this relationship with their metadata and the composition and verification of their internal cross consistency. The adherence of the behavior of the value attribute to the expected by the modeling included in their metadata, allows checking the quality for the intended use of information.

SECTION V.
Future Works

Other SDWF source systems have to be also analyzed, as well as their data, using the methodology here proposed.

A study of the evolution of the degree of data quality according to [8] is also being considered as a continuation of the work. The approach of [8] which appraises the quality of data is a function of the use of these data, so the data with higher quality are those more frequently used. Based on control systems with feedback, it mentions that the best way to improve the quality of information is to expand and improve its use.

One justification is that by repeatedly using the data, the user provides a feedback in the information system, incorporating their assessment and knowledge of data evolution, or their familiarity with the data.

Using this approach, will be verified with the operators of the transmission systems what are the information they use most often. With this information, these data will be checked for quality in relation to the other data.

Additionally, the evolution of the quality of these data will be checked along the period of their utilization.

FOOTNOTES

No Data Available

REFERENCES

1. A. T. S. CALAZANS, "Qualidade da Informação: conceitos e aplicações," TransInformação, vol. 20(1), pp. 29-45, jan/fev. 2008.
Show Context

2. F. FAVARETTO, "Melhoria da Qualidade da Informação no Controle da Produção: Estudo Exploratório Utilizando Data Warehouse," Produção, vol. 17(2), pp. 343-353, mai/ago. 2007.
Show Context

3. T. C. REDMAN, "The impact of poor data quality on typical enterprise," Communications of the ACM, v. 41(2), pp. 79-82, feb. 1998.
Quick Abstract | Show Context | Access at ACM

4. G. C. M. AMARAL, "AQUAWARE:Um Ambiente de Suporte à Qualidade de Dados em Data Warehouse," Dissertação de Mestrado, Instituto de Informática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, 2003.
Show Context

5. L. PIPINO, Y. LEE, R. WANG "Data Quality Assesment," Communications of the ACM, vol. 45(4), pp. 211-218, abr. 2002.
Quick Abstract | Show Context | Access at ACM

6. R. R. OLETO, "Percepção da Qualidade da Informação," Ci. Inf, v. 35(1), pp. 57-62, jan/abr. 2006.
Show Context

7. A. M. P. M. da COSTA. "A Gestão da Qualidade dos Dados em Ambientes de Data Warehousing na Prossecução da Excelência da Informação," Dissertação de Mestrado. Escola de Engenharia. Departamento de Informática. Universidade do Minho, 2006.
Show Context

8. K. ORR "Data Quality and Systems Theory," Communications of the ACM, vol. 41(2), pp. 66-71, fev. 1998.
Quick Abstract | Show Context | Access at ACM

9. D. BALLOU e G. TAYI, "Examining Data Quality," Communications of the ACM, vol 41(2), pp. 54-57, feb. 1998.
Quick Abstract | Show Context | Access at ACM

AUTHORS

M. R. Bastos
More by this Author: IEEE Xplore

J. S. C. Martini
More by this Author: IEEE Xplore

J. R. A. Junior
More by this Author: IEEE Xplore

S. Viana
More by this Author: IEEE Xplore

CITED BY

None

KEYWORDS

IEEE Keywords

Context, Data structures, Data warehouses, Databases, Maintenance engineering, Measurement

INSPEC: Controlled Indexing

energy management systems, information systems, power control, power supply quality

INSPEC: Non-Controlled Indexing

computational modules, data integration, data retrieval, data storage, electric power control, electric power supervision, electrical system management, events repository system, information system, large industrial process, quality aspects, quality measurement

Authors Keywords

data quality, data repository, information management, information quality, information quality perception, metadata, quality metrics

CORRECTIONS

None