How to use this CD-ROM

You need

1) a WEB Browser. In most computers, a WEB Browser already exists. If you have any problem, contact us: cscc@worldses.org

2) Acrobat Reader (See the previous page of this cover).

FAQ (Frequently Asked Questions)

When I click on a link to a PDF file, nothing appears !

Probably your version of Internet Explorer, Netscape or Acrobat Reader is not recent enough, please try installing the ones provided in the CD-ROM. They are recent enough to work properly.

I have some trouble with printing (only one page printed, some additional text printed sometimes)

It s probably because you are printing from the Browser menu or icon bar. Try printing from Acrobat Reader icon bar. There is a small print button on Acrobat Reader icon bar which is located under the Browser icon bar.

How can I make zoom in the PDF files?

Click the appropriate button down and left or the buttons with the predefined standard zoom, up and right (below the Browser).

How can I make "copy" - "paste" in PDF files?

First Click the button with the letters abc, drag the mouse and then click at the "copy" button (up and left - just left from the button with the "palm"). After that, you can make "paste" in every Windows application.

How can I find something in a text?

In the HTML files, by pressing CONTROL+F. In the PDF files (Acrobat Reader), by pressing the button with the binoculars. You can also read the PDF files of this CD-ROM by using only the Acrobat Reader (without any Browser).

ISBN: 960-8052-19-X

ISBN of Vol.2 CSCC'99: 960-8052-01-7

4th World Multi-Conference on:
Circuits, Systems, Communications and Computers
(CSCC 2000)

2nd International Conference on:
Mathematics and Computers in Physics
(MCP 2000)

2nd International Conference on:
Mathematics and Computers in Mechanical Engineering
(MCME 2000)

IMCS

STIR PALACE
VOULIAGMENI
ATHENS
GREECE.
JULY 9-16
2000

4th World Multi-Conference on:
Circuits, Systems, Communications and Computers
(CSCC 2000)
(also containing Late Papers of CSCC'99)

2nd International Conference on:
Mathematics and Computers in Physics
(MCP 2000)

2nd International Conference on:
Mathematics and Computers in Mechanical Engineering
(MCME 2000)

ISBN: 960-8052-19-X
ISBN of Vol.2 CSCC'99: 960-8052-01-7

In order to access the documents a WEB Browser (Microsoft Internet Explorer or Netscape Communicator) and the Acrobat Reader software are required.
Some versions of the Acrobat Reader software are included for your convenience, which can be run on several platforms. In this CD-ROM the **Acrobat_Reader** directory contains the **latest versions (JUNE 17, 2000)** of the Acrobat Reader software for the following platforms.
1) Windows 95/98 2) Windows NT 3) Windows 3.1 4) Windows 2000 5) Solaris 6) Sun OS

Attention: Don't install the **Acrobat_Reader** , if you have already a newer version of it.
If you want to set up the ACROBAT READER program for other OPERATING SYSTEMS (Macintosh PPC, Macintosh 68K, IBM AIX, SGI IRIX, HP UX, Digital Unix, etc), please, visit http://www.adobe.com You can download the Installation programs from
http://www.adobe.com/products/acrobat/readstep2.html
If there is any problem installing them, we advise you to contact Adobe Acrobat Inc. Ph: 1-900-555-2276, Mon-Fri: 6AM - 2PM (Pacific Time) or contact us: cscc@worldses.org
If you want to set up the ACROBAT READER program for other languages (for example Chinese, Arabic, French, Spanish, Portuguese, German, Italian, Japanese, Korean, Swedish, Hebrew), please, visit also http://www.adobe.com/products/acrobat/readstep2.html
Now, you can receive free weekly Adobe newsletters, plus the latest technical information, via e-mail. Please, visit: http://www.adobe.com/services/newsletter/subscribe.html and subscribe.

Make sure you install the Acrobat Reader with Search if you download from the Web Site. Search routine helps to locate any paper based on a keyword of your choice.

Data Quality and Database Design[1]

MARIO PIATTINI, ISMAEL CABALLERO,
MARCELA GENERO, CORAL CALERO.
Grupo Alarcos. Escuela Superior de Informática.
Universidad de Castilla-La Mancha.
Ronda de Calatrava 7 13071 Ciudad Real.
ESPAÑA.
mpiattin@inf-cr.uclm.es http://alarcos.inf-cr.uclm.es

Abstract: - Both products and services must satisfy customers' requirements. Information Systems and their Databases are the main support for organizations to collect, store, and retrieval these requirement data. If any of these operations are badly executed or not made on the right data, they will not produce useful results, and our aim will not get satisfied. That is the reason for which we are interested in data quality. This paper deals about what data quality is, which are the most important dimension of data quality and how we can design quality databases.
Key-Words: - data quality, quality dimensions.

1 Introduction

Nowadays, most companies are facing a severe problem of data pollution, i.e. they have too much data at their disposal, mainly due to three reasons:

- Data can be captured in a very easy and inexpensive way, due to recent improvements and diffusion of data entry technology: barcodes, OCR (Optic Character Recognisers), customer cards, credit cards...). Besides, lots of data can be directly obtained from the Internet.
- Uncontrolled data redundancy: Due to their daily functioning, information systems grow in a disorderly and unplanned way, and in many cases companies do not have an information architecture.
- Existence of large quantities of historical 'expired' data, which no longer serve to carry out any kind of process nor to obtain any relevant information.

As [6] points out, this can be paralleled to a biological process, so that data that are not used tend to become atrophied. This pollution can have serious consequences; thus, for example, [1] firms that up to a half of the total cost of a data warehouse implementation may be caused by a poor data quality. The Gartner Group has warned that poor data quality has been one of the most frequent failure causes in reengineering projects. Therefore, it is necessary for the correct operation of the company Information System (IS) to address the problem of data quality, so that data becomes actual information and knowledge. Companies must manage information as an important product, capitalising on knowledge as a main asset in order to survive and prosper in the age of digital economy ([3]). By improving information quality, we will improve both client and personnel satisfaction, which will further contribute to the improvement of the whole company.

Unfortunately, research on quality has focused until very recently on software quality, neglecting data quality ([9]). Even in the case of traditional database design, quality has not been explicitly incorporated ([14]). Although databases have not traditionally focused on questions of quality, many of the tools and techniques developed (integrity constraints, normalisation theory, transaction management, etc.) have influenced quality. We think it is time to consider information quality as a main objective, rather than as a by-product of the process of database creation and development.

Broadly speaking, two different aspects should be kept in mind regarding quality information: quality of the data base as a whole and quality of data presentation. In fact, it is very important that the data reflect the real world in a correct way, that is, that they are precise; moreover, they must be easy to understand. Regarding data base quality as a whole, it depends on three 'qualities': DBMS (Data Base Management System)

1 This research has been done within the framework of the CALIDAT Project, developed by Cronos Ibérica, S.A. in collaboration with the Universidad de Castilla-La Mancha, supported by the Consejería de Educación y Cultura de la Comunidad de Madrid (Ref: 09/0013/1999).

quality, data model (both conceptual and logical) quality and data quality. In this paper we will focus on the most prominent features of the data; regarding data model quality, the interested reader can consult [7].

2 Dimensions of information quality

As we know, quality is a relative concept, as far as it is in the eyes of the beholder; for this reason, we can consider quality as a multidimensional concept, subject to restrictions and limitations ([4]). In recent years, various authors have proposed different dimensions for data quality:

- [8] groups quality dimensions into three self-explanatory categories:
  - Quality Dimensions of a Conceptual View
    - Content: Relevance of the data, obtainability of values, clarity of definition.
    - Scope: Comprehensiveness and essentialness
    - Level of Detail: Granularity of attributes and precision of domains.
    - Composition: Naturalness, identifiability, Homogeneity, Minimal unnecessary Redundancy.
    - View Consistency: Structural and semantic consistencies.
    - Reaction to Change: Flexibility and robustness.
  - Quality Dimensions of Data Values
    - Accuracy
    - Completeness
    - Currency
    - Value Consistency
  - Quality Dimensions of Data Representation
    - Appropriateness
    - Interpretability
    - Portability
    - Format Precision
    - Format Flexibility
    - Ability to represent null values
    - Efficient usage of Recording Media
    - Representation Consistency
- [2] emphasises two topics related to data quality:
  - Inherent quality, that is, data accuracy, the degree to which the data reflect the objects from the real world they represent; this includes: conformity with the definition, completion of values, validity or conformity with the company rules, accuracy of sources, accuracy of reality, lack of duplication, accessibility.
  - Pragmatic quality, i.e. the degree to which the data allows the knowledge workers to satisfy the company objectives in an efficient and accurate way.
- [13] analyse some of the causes of poor data quality due to design deficiencies from an ontological perspective, identifying four quality dimensions:
  - Data Quality
  - Nature of deficiency
  - Completion
  - Improper representation

As these authors indicate, the aim is that each state in the real world will unequivocally correspond to one system state. If the unequivocal relationship is not verified, or the expected results are not obtained when operating with the data, a data deficiency is produced.

- [11] identify several dimensions group by four categories:
  - Intrinsic: precision, objectivity, credibility, reputation
  - Accessibility: accessibility, access security
  - Contextual: relevance, added value, opportunity, completion, data quantity
  - Representational: interpretability, comprehension facility, concise representation, consistent representation.

3 Data base design and data quality

[5] suggest three different strategies in order to improve the intrinsic quality of data bases:

- Building richer semantic models that reflect reality more accurately.

- Reinforcing databases by introducing a higher number of constraints, in order to identify and discriminate problematic data and link them to the appropriate applications.
- Restricting the use of data to predefined processes, preventing them from being modified by other processes so that they cannot be accidentally deleted.

Although these strategies allow a higher degree of data quality, they are not enough by themselves, since an adequate base is needed in order to manage quality dimensions ([16]). Unfortunately, there are very few proposals that consider data quality as a fundamental factor in the design process. In this sense, [14, 15] are an exception to this rule. They propose a method that is intended to complement the traditional design methodologies of database design, see figure 1 at the end of the paper.

In the first step, see figure 1, apart from creating a conceptual scheme, e.g. using an entity/ relationship model, quality requisites and candidate attributes should be identified, determining thereafter the 'quality parameter view', so that each element within the conceptual schema can be linked to a quality parameter. E.g., in an 'academic' data base, the attribute 'exam mark" can be linked to precision and timeliness. Later on, subjective parameters are objectivated through the addition of labels to the attributes in the conceptual scheme (source, in order to know the degree of accuracy, and date, in order to know the timeliness, of exam marks).

Moreover, we can also propose an extension of relational databases with indicators that allow the assignment of these objective and subjective parameters to the quality of the values within the data base ([15]).

4 Conclusions and future research

We can affirm that, if product and service quality has become a decisive factor of business success in recent years, information quality will receive a preferential role in the next decade.

If we actually consider that information is the most important business asset, one of the first aims of IT professionals should consist in ensuring its quality. We have presented some recent proposals regarding information quality, but further research is needed on the degree of quality attached to other processes linked to information: modelling, data gathering and loading, and data presentation.

On the one hand, companies will have to define a quality policies (see, for example, [8]) that defines the obligations of each function in order to ensure data quality in all its dimensions; on the other hand, they will have to implement a process in order to evaluate the quality of the information at their disposal. There are several proposals regarding information quality evaluation; English's TQdM (Total Quality data Management) can be highlighted.

A decisive aspect regarding evaluation has to do with the definition of relevant metrics, that will allow an actual analysis and improvement of quality. In [3], three types of metrics are proposed: subjective metrics (based on the data users' judgement), objective metrics that are independent of the application (such as correction) and objective metrics belonging to the application (i.e., that are specific to a given domain). Besides, the actual value of information (either produced by operational systems or used to assist decision taking) should be measured.

References

[1] Celko J., Don`t Warehouse Dirty Data. *Datamation,* 15 October, 1995, pp. 42-52.

[2] English, L. *Improving Data Warehouse and Business Information Quality*. John Wiley & Sons, Inc., 1999.

[3] Huang, K-T., Lee, Y.W. and Wang, R.Y. *Quality Information and Knowledge.* Prentice Hall, Upper Saddle River, 1999

[4] Jones, C. *Software Quality. Analysis and Guidelines for Success*. London: International Thomson Computer Press, 1997.

[5] Orman, L., Storey, V. Y Wang, R. Systems Approaches to Improving Data Quality. TDQM-94-05, August 1994. Available on http://web.mit.edu/tdqm/www/papers/94/94-05.html

[6] Orr, KData Quality and System Theory. *Communications of the ACM,* 41 (2), 1998, pp. 66-71.

[7] Piattini, M., Genero, M., Calero, C., Ruiz, F. and Polo, M. Database quality. In: *Advanced Databases: Technology and Design*. Piattini, M. and Diaz, O. (eds.). London, Artech House, 2000

[8] Redman, T. C. *Data Quality for the Information Age*. Artech House, Boston, 1996.
[9] Sneed, H.M. and Foshag, O. Measuring Legacy Database Structures. *Proc. of The European Software Measurement Conference FESMA '98*, Coombes, Hooft and Peeters (eds.), 1998, pp. 199-210.
[10] Storey, V. C. and Wang, R. Modeling Quality Requirements in Conceptual Database Design. TDQM-94-02, May 1994.
[11] Strong, D.M., Lee, Y.W. and Wang, R.Y. Data Quality in Context. *Communications of the ACM*, Vol. 40, No. 5, 1997, pp. 103-110.
[12] Strong, D.M., Lee, Y.W. and Wang, R.Y. 10 Potholes in the Road to Information Quality. *IEEE Computer*, 1997, pp. 38-46.
[13] Wand, Y. and Wang, R.Y. (1996). Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*, Vol. 39 (11), 1996, pp.86-95.
[14] Wang, R. Y., Kon, H. B. and Madnick, S. E. (1993). Data Quality Requirements Analysis and Modeling. *Proc. of the 9th International Conference on Data Engineering, IEEE Computer Society*, 1993, pp. 670-677.
[15] Wang, R.Y., Reddy, M.P. and Kon, H.B. Toward quality data: An attribute-based approach. *Decision Support Systems*, Vol. 13, 1995, pp. 349-372.

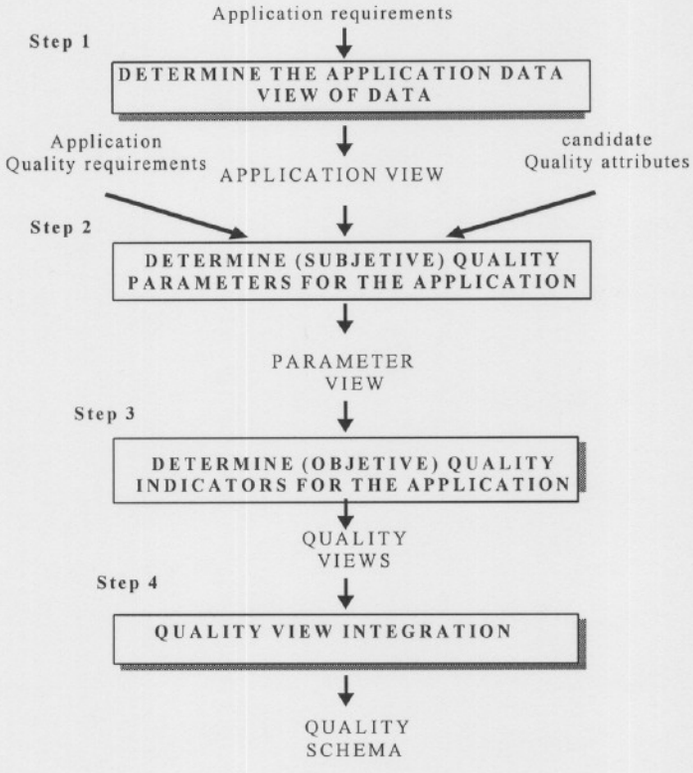

Fig. 1. Quality in database design ([14])