Data representation factors and dimensions from the quality function deployment (QFD) perspective

Maria Pinto

Department of Information Science, University of Granada, Spain

Received 19 June 2005
Revised 5 September 2005

Abstract.

In order to optimize access to the increasing amount of information, a classic solution has been data representation. The aim of this research is to uncover and systematize the factors and dimensions involved in the data representation issue and more exactly in the planning and design of the information products (IP) and their previous representation processes (RP). QFD (quality function deployment) is a planning tool based on user needs and expectations – quality functions – allowing the planning and design of IPs and RPs. A series of linked deployments provides the implied factors and dimensions: IP planning factors – representing, relating, filtering and seeking relevant information; IP design dimensions – relevance, format, comprehensiveness, consistency, accuracy and currentness; RP planning factors – comprehension, synthesizing, structuring and selecting; and RP design dimensions – human resources, computers and tangibles. By means of these deployments, the analysis of the factors and dimensions and their corresponding relationships provides an excellent picture for the quality planning and design of information products and representation processes.

Keywords: abstracting quality; data quality; data representation; information quality; quality function deployment; representation processes; total data quality management

1. Introduction: research purpose and contribution

In spite of the vast amount of research on database quality carried out over the last decades, there is a need for a stronger emphasis on data investigation. The enormous growth in databases of all sizes and designs creates the necessity for better methods to access and analyse the data [1]. There is clear evidence that error rates in some existing databases are unacceptably high [2]. Inappropriate, ambiguous or deliberately fraudulent, biased or non-objective, incomplete and out-of-date data are the most extended problems related to data quality [3]. Apart from this evident lack of data perfection and objectivity within many databases (DBs), one has to be conscious that the consumer’s needs are not given once and for all, but are subject to change [4]. The concern here is data usefulness, a quality concept founded on user needs [5–7]. Yet the ample and fuzzy concept of data is restricted to the referential data, or data representing the original sources and that consequently refer or direct users to source documents [8]. In fact referential data are information products (IPs) derived from the corresponding
information manufacturing or representation processes (RPs). Like physical products, IPs can be grouped according to similar characteristics and common data inputs permitting the group to be managed as a whole [9]. The quality paradigm is a subjective construct [10] with an ergonomic origin [11] and interrelated dimensions [12]. Regarding the fitness for purpose and the multidimensional and holistic features of the quality paradigm, what we need is a systematic view of IP and RP quality, chiefly rooted in users.

However, the procedures of systematic quality planning and design of documentary products are a complex matter [13]. In an industry that is essentially a service, and whose primary currency is intangible information, quality is not only difficult to define; it is difficult to quantify [14]. Data quality research is a complicated interdisciplinary field spanning diverse disciplines such as management, computer science and psychology [15]. A set of works on DB evaluation has been developed with special emphasis on content quality. Various research works stand out especially, such as those of the Finnish Society for Information Services [16], the Southern California Online Users Group (www.scougweb.org/) and Peter Jacso [17]. Also the Centre for Information Quality Management (CIQM) is a service of Information Automation Ltd, run on behalf of the UK e-Information Group (formerly UKOLUG) and CILIP (initially for the Library Association) [14]. There seems to be a pragmatic consensus concerning data quality’s scope and key elements [17], and at the core of all aspects is content, the information itself as created by the author, secondary publisher or database producer [18,19]. Nonetheless, the design of information products usually involves modelling the information environment and the information mission for which the database is being implemented [1]. From this quality perspective, IPs and RPs are addressed in a triple way: they depend on users, with their individual and collective expectations and perceptions of performance; they are conditioned by the context (the database system and its specific mission); and they have to be based on raw data content.

From the above one can deduce that discrete quality-oriented practices are unlikely to be effective unless they are an integral part of an organizational system for quality improvement [20]. Within this systemic line of investigation, the specific goal of this paper is to uncover and systematize the drivers of the raw data representation issue, that is, the key factors and dimensions to be taken into account at the time of conceiving IPs and RPs from the quality viewpoint. The terms factor and dimension are used here with the meaning of something that is in the state of contributing causally to a wanted result. While everything that drives the planning of IPs and RPs is regarded as a representation factor, what drives their design is considered to be a representation dimension. Though the planning and design concepts have scarcely been introduced within the information-documentation domain, when conceiving a quality product they are the two basic and unavoidable milestones. If the core of this research is the planning/design – factors/dimensions – of the representation issue from the quality perspective, a good approach to the topic should be the Quality Function Deployment (QFD) methodology, a successful tool in the industrial sector that may provide good results in the information and documentation domain. With the feeling of being at the starting point of a new path – and probably a new race – our purpose here is limited to the conception – planning/design – of the main factors and dimensions involved in the raw data representation issue. Our choice is the wider concept of data representation regarding all kinds of data, in comparison with that of metadata, a very similar idea but restricted to the electronic field.

This paper is structured as follows: we present a first section on the theoretical foundations of the work and its conceptual and methodological framework, and then we posit our research design, offering a description of the data gathering and analysis, the user population and the questionnaires included. The research deployment and outcomes section constitutes the core of the paper, explaining the QFD basic steps and showing the research results. The last section contains some conclusions.

2. Conceptual and methodological framework:

2.1. QFD

The conceptual and methodological framework is person-oriented and starts from an empirical approach that, after collecting information from consumers, will help us obtain a comprehensive framework of data quality from data consumers’ perspectives [21,22]. The base is Total Quality Management (TQM), a systematic and holistic approach to the management of organizations whose guiding principles focus on customer satisfaction, organize work as a process, measure results, recognize that people are key elements to success, and foster a culture of continuous improvement [23]. Fundamental precepts of TQM are that:
Data representation and QFD

- organizations should be viewed as systems of interlinked processes;
- quality problems cannot be addressed by patchwork solutions; and
- management should focus on the creation and perpetuation of an organizational system geared toward the achievement of superior quality performance [20].

Within this line of research the Total Data Quality Management (TDQM) Cycle, an adaptation of TQM principles to the context of data, consists of defining, measuring, analysing and improving data quality through multiple, continuous improvement cycles [24]. The TDQM Research Program was conceived to deliver high-quality information products to information consumers [22, 25–27]. Organizations need outside assistance to make the transition from a largely internal to a more balanced internal/external focus on customer value, combining perceptions, preferences and evaluations [28].

An important branch of TQM is Quality Function Deployment (QFD), a philosophy, but also a methodology that drives the transformation of customers’ demands into quality characteristics. The aim is to develop a design quality for the finished product by systematically deploying the relationships between the demands and the characteristics, starting with the quality of each functional component and extending the deployment to the quality of each part and process. The overall quality of the product will be formed through this network of relationships [29]. Characterized by customer orientation, cross-functional management and process rather than product orientation [30], its principal purpose is to ensure that the needs and wants of customers drive the product generation process. It employs visual planning matrices and other quantitative means to integrate customer requirements, design requirements, target values, and competitive performance [31]. Based on the idea that the customer is the only voice worth listening to [32], QFD adapts technology to people and can be used to develop ergonomic or usable products [11]. Customer requirements and their relationships with planning and design characteristics are the driving forces of QFD methodology [30].

The first and key phase of the QFD process consists of translating the categorized and ranked customer requirements, that is to say the quality functions (QF), into technically quantifiable terms. This first key stage, also known as product concept development – concept engineering – is a detailed and structured process for enhancing the initial stages of QFD, integrating customer-driven requirements into structured design activities [33]. In this meaningful phase we take each quality function QF (named WHAT by the QFD practitioners) separately and ask ourselves how we can get it. The answer is a matching list of planning requirements (named HOW). For a better view of the process, a graphic deployment of the QFs is needed, and so we plot all the HOWs at right angles to the WHATs, setting up a matrix in which the relation of each QF to each planning requirement is represented by a separate square (see Figure 1). We can then use this matrix to determine whether there is any relation between each QF and each planning requirement, and, if so, how strong the relation is. Conventionally, within the matrices the rows correspond to the WHATs and the columns to the HOWs. Their relationships are represented by means of symbols specifying the degree of relationship: a filled-in circle shows a strong relationship between the corresponding row and column; an empty circle shows a moderate relationship; and a triangle a weak one. Once we have developed the relationships matrix, identifying the hierarchy is a relatively easy task. The numerical calculation of the absolute importance for each technical descriptor (HOW) is the product of the cell value – (1, 3, 9) depending on the type of relationship (weak, moderate, strong) – and the customer importance rating (WHAT). Numbers are then added up in their respective columns to determine the importance for each technical descriptor (HOW MUCH). As the various HOWs are elements of the same system, and are rarely independent of one another, it is interesting to analyse the possible correlation that may exist between them, and, wherever such correlations exist, whether the parameters reinforce or counteract each other. A ‘roof’ over the HOW – the correlation matrix – is built, and this is why this first deployment is also known as the House of Quality. On the basis of the first matrix – the House of Quality – in Figure 1, which is a common referent for later stages, we can define a product concept as well as target figures for the process behind it. The HOWs – IP planning factors – and their target figures from the first matrix are the WHATs in the second stage matrix in Figure 2. The HOWs from the second matrix – IP design dimensions – are the WHATs for the third stage matrix in Figure 3. The HOWs from the third matrix – RP planning factors – are the WHATs for the last matrix in Figure 4. An alternative technique is Simplified QFD (S-QFD), which focuses on the major product and process constraints, rather than on an exhaustive listing of all requirements and issues [34]. The S-QFD version, whose only specificity is that the
number of matrix elements has been simplified, constitutes the basis of the remainder of this paper.

3. Research design

For the user-centred approach to the data representation topic promoted in this research, the first need that arises concerns a knowledge of the user population for whom the data are intended – who these people are, and above all the nature of their needs, expectations and perceptions. Once we have determined the user population, a knowledge of the user profile with its needs, perceptions and expectations is possible. User profiles provide an understanding of the users’ literacy, experience, work situation and information needs, eliminating the assumption that we know what people want, when they want it, and how they want it delivered. This prevents us from developing products and services that satisfy our needs rather than those of our users [35]. Later, QFD methodology will provide the leap from the problem-space of these user demands to the solution-space of the data representation factors and dimensions.

3.1. Data gathering

The process of data gathering has two basic stages: definition of the users and collection of questionnaires addressed to the user population.

3.1.1. User definition. Much of the scepticism about user studies is related to the lack of adequate theories guiding this research. It has been explicitly formulated that such a theory would appear when enough empirical data had been collected [36]. With the aim of gathering enough empirical starting information, a double universe of people was defined. On the one hand, the managers, abstractors and other individuals within the Spanish abstracting documentation centres; on the other, the researchers who are users of abstract journals. For the first group, a preliminary map of candidate documentation centres (51) was established. Because almost half of the centres did not elaborate authentic abstracts – providing only an author-abstract – the research area was reduced to 27 public and private centres of the most varied kinds: social sciences, humanities, education, social affairs, communications, statistics, and safety, for example. For the definitive selection, non-random intentional sampling was used with the aim of ensuring geographical representativeness; a total of 20 Spanish documentation centres remained. For the second group, researchers and other users of abstract journals, attention was centred on the ‘Plan Andaluz de Investigación’ (PAI) database [37] within the domains of the PAI’s social, economic, and legal sciences that collect the combined research work (produced by teachers and academics) of all the Andalusian universities. Once we had obtained and analysed the candidate population (2427 researchers), a random sample of 525 was selected in order to ensure the representativeness of the domains under study.

3.1.2. Questionnaires collection. No matter what the researcher’s experience on the abstract/abstracting issue, it was necessary to provide scientific evidence of the empirical starting point toward the uncovering of the main factors and dimensions of this issue. The choice of the abstracts is not fortuitous, considering that these are the most complex and representative IPs of referential DBs. A first questionnaire with 29 items was sent to abstractors from documentation centres and other people concerned, with a good percentage of closed questions on:

1. organization of the abstract service – status, human resources, financial resources, attributions;
2. abstract production – processes, typology, design, forms of production, abstract structure, extent; and
3. policies and strategies of the service – coverage, format, instructions, recommendations and regulations, users.

In any case the basis for the questionnaire design was to find out what abstract functions are considered more relevant by the respondents. The second questionnaire – addressed to researchers and DB users, particularly of abstract journals, was designed around 10 items of closed questions related to the desired abstract functions. While in the first case the number of responses received was 41 (66%), in the second one it was 298 (57%).

3.2. Data analysis

The research was now at the stage of defining the quality functions to be deployed. A prior condition is to reduce the number of requirements to manageable proportions [38]. The classification of the questionnaire answers drives the categorization of the data representation issue. Thus, the three most important QFs for the IPs to be accomplished were selected: accessing relevant data (55), gathering relevant data (30), and
Data representation and QFD

Table 1
Quality functions from the data analysis

<table>
<thead>
<tr>
<th>Quality function</th>
<th>Relative importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessing relevant data</td>
<td>55</td>
</tr>
<tr>
<td>Gathering relevant data</td>
<td>30</td>
</tr>
<tr>
<td>Updating relevant data</td>
<td>15</td>
</tr>
</tbody>
</table>

*updating* relevant data (15). The numbers in brackets are percentages showing the relative importance of each QF (see Table 1), being derived from the absolute numbers of responses of the sample. These items came about directly, without interpretation, because they were part of a collection of closed answers. The reduction of the number of quality functions down to three is a research restriction with no other goal than the control of the number of involved variables, which otherwise could swell to unmanageable proportions.

Once we had defined the data representation QF profile, the research reached the starting point of the QFD process, whose goal is to transform QFs into IP and RP factors and dimensions.

4. Research deployment and outcomes

The three selected QFs are the starting point of the QFD process undertaken. For this to be developed, the QFD-capture software by ITI used throughout the research has been extremely useful. But these functions and the resultant quality factors and dimensions deployed along this research are neither software-dependent nor prescribed by the QFD methodology. They are rational answers to the stated questions, according to the well-known principles of problem solving and decision making. Even with the background of a specific methodology and the help of the parallel software, we face a difficult and creative task.

4.1. IP planning factors

IP planning is the first step of the S-QFD research model. The three-sided mission of DBs when reflecting simplified and categorized customer requirements (*accessing, gathering and updating*) has been deployed into four planning requirements. This is a creative answering process that may be collectively performed. For the deployment of the three *quality functions* a poll among our higher level students at Granada’s Faculty of Documentation was helpful at the time of taking the decision of choosing *representing*, *relating*, *filtering* and *seeking* relevant original sources as the four functional forces that drive the planning of IPs. Taking into account that the number of deployed functions has to be a minimum, we obtained the four IP-planning factors shown in section How of Figure 1. The strength of the relationships among the Whats and the Hows was also a research decision. In fact, all factors and dimensions for the remainder of the paper have surfaced in this way.

The top quality roof is also derived from a reasoning process. Each factor is defined in detail in the following sub-sections, each named according to the factor discussed. The various factors of the model are discussed using theoretical reasoning and/or empirical evidence.

4.1.1. Representing. From the survey and our own reflections on the issue it was decided that the factor of representing would be strongly (9) linked to accessing (55) and moderately (3) to gathering (30) quality functions. From the result of the sum of \((55 \times 9)\) and \((30 \times 3)\), slightly diverted by a comparative analysis
performed by the software, a relative importance among IP-planning factors of 54% is derived.

Representing is not a straightforward process. How best to characterize the physical and intellectual attributes of books and other physical media was debated well before the advent of computerization, and the invention and popularization of non-physical information carriers have neither simplified the questions nor resolved the debates [39]. Several types of representations have played a prominent role, among which the mental-space model, the featural model, and the structured-representations model stand out. The mental-space model thinks of the representing world as a continuous and geometric space relative to a set of axes. The featural model assumes that mental representations consist of lists of symbols. Features are discrete elements that can be assessed, reported and used, in contrast to spatial dimensions, which are continuous. The third model, related to structured representations, is the most complete form of representation, highlighting the importance of the links that exist among the elements. Semantic networks are simplified variants.

Because of political, religious, cultural, gender and language differences, knowledge representation and organization systems have no accepted standards of representational accuracy [39]. It will be vital to investigate the linguistic, communicative, and organizational aspects of representation from a multiplicity of socio-cognitive perspectives and within the full range of discourse domains and knowledge communities [40]. Though numerous methods have been developed, the latest are based on document structure, text as a discourse unit, and natural language generation and processing [41]. In any case, external representations are not simply inputs and stimuli to the internal mind. For many tasks, external representations are so intrinsic to the tasks that they actually guide, constrain, and even determine the pattern of cognitive behaviour and the way the mind functions [42]. The activity of representing and organizing knowledge can never be viewed as an end in itself. Rather, it is the ongoing process of constructing and reconstructing effective infrastructures that will support productive activity and encourage dialogue, both internally within a knowledge domain and externally across the boundaries of competing or conflicting domains [43]. The challenge is to develop methods for integrating different representational assumptions into broader models of cognitive processing.

Regarding the results of the deployment, representing (54%) is the main factor at the time of IP planning, being linked to some IP-design dimensions of the next deployment: strongly to consistency, comprehensiveness, accuracy and relevance, and moderately to format (Figure 2). Therefore, its impact on IP design is significant.

4.1.2. Relating. The relationships among the IPs of a database smooth the path toward the accessing and gathering of their corresponding original sources, and so the relating factor has been considered moderately linked to the accessing and weakly to the gathering quality functions. Therefore, its relative importance among the studied IP-planning factors is 18%.

The IP-planning factor of relating is used here in the sense of associating, linking, or making a logical or causal connection among the different data representations. The fact is that the triangle relation–association–connection leads to the concept of interaction. Information and communications technologies (ICTs) are fundamentally changing the nature of the learning experience that at its best allows and encourages new forms of interaction. The term ‘networked learning’ describes the range of educational approaches which are exploiting the new technologies [12]. It would seem likely that the importance of skills in selecting, assessing, using and processing information will increase within networked environments. The development and refinement of such skills will form an essential underpinning for the successful adoption of networked learning. A common ground to all hypertext systems, regardless of their location on the association/connection continuum, is the issue of non-linearity of access to information [44]. Hypertext is an approach to information management in which data are stored in a network of nodes and links, connoting a technique for organizing textual information in a complex, non-linear way to facilitate the rapid exploration of large bodies of knowledge.

Functionally speaking, the importance of the relating factor obtained through the deployment is considerable (18%), and this is even more understandable taking into account its great reliance on some dimensions in the next stage of IP design: a strong reliance on format and weak on relevance (Figure 2).

4.1.3. Filtering. In order to gather and update the original relevant data, there is an unavoidable need to filter the information contained within the data. For this reason, in our first IP-planning deployment, the filtering factor has been considered strongly linked to the gathering and weakly linked to the updating quality functions. From the result of the sum of (30 × 9) and
Data representation and QFD

...a relative importance among IP-planning factors of 16% is derived.

Electronic message distribution systems can contribute to information overload, when the information is received at such a fast rate that it cannot be assimilated [45]. In order to carry out the removal of this excess of information a filter is required, and to this end a model must be built to fit the user’s interests – a filtering profile, based on the topics of interest to the user. A formal model of the decision process used in deciding whether to examine a message in an organizational setting has been developed [46]. The model, based on economic and statistical decision theory, and making use of work done on information retrieval and automatic indexing systems, suggests a method whereby unexamined messages are ranked [47] from those most likely to be of interest, or ‘relevant’ to our needs, to those least likely to be relevant. Filtering criteria related to the subject matter can sometimes be expressed most effectively in a negative way, by pointing to subject attributes that should not be present in retrieved documents [48].

The relative importance obtained for the IP-planning factor of filtering is 16%. Within the next deployment, it will be strongly related to the IP-design dimension of relevance (Figure 2).

4.1.4. Seeking. When one wants the relevant data to be updated and gathered, there is a need to seek the information. In our proposed first deployment, the IP-planning factor of seeking has been strongly linked to the updating and weakly to the gathering quality functions. As a result, its relative importance among IP-planning factors is 12%.

Information seeking is one field of recent and promising research, and although there are an increasing number of models of information-seeking behaviour, our understanding of this behaviour is still developing [49]. In any case, information seeking includes not only the initial location of information, but also its subsequent management and relocation: our concern is not only with the initial retrieval of information, but also with the organization and use of information throughout the wider constructive task [50]. In all the information-seeking activities, users look for information of potential interest, making judgments of information quality and cognitive authority, and then interpret the information [7]. Information seeking is a process of construction which involves the whole experience of the person, feelings as well as thoughts and actions [51]. The search process, particularly during the entry and orientation phases, is subtler and more complex, on several grounds, than current models assume. A model representing the user’s sense-making process of information seeking ought to incorporate three realms of activity: physical, actual actions taken; affective, feelings experienced; and cognitive thoughts concerning both process and content. A person moves from the initial state of information need to the goal state of resolution by a series of choices made through a complex interplay within these three realms [52]. Also, collaborative information seeking has begun to emerge as an important field of study [53]. Information retrieval (IR) and artificial intelligence (AI) are the two main roads for seeking information. While IR methods include Boolean, vector space and probabilistic models, more advanced models, based on a combination of logic and uncertainty theories, have been proposed for handling multimedia-structured data. AI methods enable intelligent information access, as is the case of ontologies, which provide explicit domain theories that can be used to make the semantics of information explicit and machine-actionable.

With a relative importance of 12%, the IP-planning factor of seeking will be strongly related to the IP-design dimension of currentness and moderately related to that of relevance (Figure 2).

4.2. IP-design dimensions

Within the second phase of this S-QFD model the IP-planning factors are deployed into IP-design dimensions. The most important at this stage is data integrity [54], and for this to be accomplished there is a need to resort to at least six basic dimensions: relevance, consistency, accuracy, comprehensiveness, format, and currentness (Figure 2).

4.2.1. Relevance. According to our S-QFD, the main IP-design dimension is relevance, because of its multiple relationships with all the factors of the former stage of IP planning: strongly linked to filtering and representing, moderately to seeking, and weakly to relating. As a result of the multiple relationships, a relative importance among the IP-design dimensions of 30% is obtained.

Since so many of the decisions we make require sifting through massive amounts of information, relevance is an important factor in making the best possible decisions [55]. Though there are distinctions to be made between system or algorithmic relevance;
is identified, can be called [60]. This view of relevance, with which this research terms of their support and contribution to a certain task sense, the relevance of data is assessed by the actors in a task- and process-oriented user construct. In this relevance has been introduced in some studies [59] as may be relevant to a different degree, multi-level relevance cannot reflect the possibility that documents phase, and in the evaluation stage [55]. As binary gathering: in the topic definition stage, in the searching ant concept to apply at various stages of information interpretation frames or hermeneutics based upon prototypical users with finite sets of criteria for relevance judgments, one may be able to evolve a better conceptual framework for designing and developing systems. However, the process of relevance judgments entails a relevance system relating the user and his need to the norms of his subject domain. Relevance is an important concept to apply at various stages of information gathering: in the topic definition stage, in the searching phase, and in the evaluation stage [55]. As binary relevance cannot reflect the possibility that documents may be relevant to a different degree, multi-level relevance has been introduced in some studies [59] as a task- and process-oriented user construct. In this sense, the relevance of data is assessed by the actors in terms of their support and contribution to a certain task [60]. This view of relevance, with which this research is identified, can be called situational.

According to our S-QFD, the relative importance of relevance (30%) is further reinforced considering the factors of the following RP-planning deployment: it is strongly related to comprehension and selecting, moderately to synthesizing, and weakly to structuring (Figure 3).

4.2.2. Consistency. The IP-design dimension of consistency has been strongly linked to the IP-planning factor of representing. Regarding that relationship, a relative importance of 16% is deduced.

Data are said to be consistent with respect to a set of data model constraints if they meets all the constraints in the set. The dimension of consistency is developed across the data elements and could be user-defined as integrity [24]. The consistency may be analysed both at the level of expressions and at the level of the concepts that the expressions represent. The former relates directly to IR techniques (matching of strings), and the latter to the deeper structures of articles, which may be consistent even if the surface is inconsistent [61]. Consistency is necessary but not sufficient for correctness [2]. Representing with quality means representing consistently, and this refers to the extent to which definitions of elements are consistent among themselves. Consistency allows predictability, helping users realize what they will and will not get in a database search [18]. Consistency means uniformity and agreement in the processing of all information units. For this to be achieved, strict compliance with rules and working instructions is necessary [19]. Consistency favours quality, as can be proved by the fact that in an excellent database all records are similarly structured [62]. Consistency has a medium relative importance of 16%, being strongly linked to the RP-planning factors of synthesizing and structuring within the next deployment (Figure 3).

4.2.3. Accuracy. The IP-design dimension of accuracy has been strongly linked to the IP-planning factor of representing. As a consequence, a relative importance of 16% is deduced.

The basic entity model [63] conceives data as triple \((e, a, v)\), where \(e\) is an entity in a conceptual model, \(a\) is an attribute of entity \(e\), and \(v\) is a value from the domain of attribute \(a\). According to this model, the accuracy of a datum refers to the degree of closeness of its value \(v\) to some value \(v'\) in the attribute domain considered correct for the entity \(e\) and the attribute \(a\) [2]. Though accuracy may be expressed as a measure of inaccuracy, quantification of data inaccuracy is a non-trivial task, if and when it is possible. At the level of

<table>
<thead>
<tr>
<th>Relationship key</th>
<th>Strong</th>
<th>Medium</th>
<th>Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

- **WHAT**
  - REPRESENTING
  - RELATING
  - FILTERING
  - SEEKING

- **Direction of Improvement**
  - 16
  - 16
  - 16
  - 30
  - 5
  - 17

- **Importance of HOW**
  - COMPREHENSIVENESS
  - CONSISTENCY
  - ACCURACY
  - RELEVANCE
  - CURRENTNESS
  - FORMAT

- **VMF IP-planning model**

Fig. 2. Second deployment. From IP-planning factors to IP-design dimensions.

16% in the set. The dimension of consistency is developed across the data elements and could be user-defined as integrity [24]. The consistency may be analysed both at the level of expressions and at the level of the concepts that the expressions represent. The former relates directly to IR techniques (matching of strings), and the latter to the deeper structures of articles, which may be consistent even if the surface is inconsistent [61]. Consistency is necessary but not sufficient for correctness [2]. Representing with quality means representing consistently, and this refers to the extent to which definitions of elements are consistent among themselves. Consistency allows predictability, helping users realize what they will and will not get in a database search [18]. Consistency means uniformity and agreement in the processing of all information units. For this to be achieved, strict compliance with rules and working instructions is necessary [19]. Consistency favours quality, as can be proved by the fact that in an excellent database all records are similarly structured [62]. Consistency has a medium relative importance of 16%, being strongly linked to the RP-planning factors of synthesizing and structuring within the next deployment (Figure 3).

4.2.3. Accuracy. The IP-design dimension of accuracy has been strongly linked to the IP-planning factor of representing. As a consequence, a relative importance of 16% is deduced.

The basic entity model [63] conceives data as triple \((e, a, v)\), where \(e\) is an entity in a conceptual model, \(a\) is an attribute of entity \(e\), and \(v\) is a value from the domain of attribute \(a\). According to this model, the accuracy of a datum refers to the degree of closeness of its value \(v\) to some value \(v'\) in the attribute domain considered correct for the entity \(e\) and the attribute \(a\) [2]. Though accuracy may be expressed as a measure of inaccuracy, quantification of data inaccuracy is a non-trivial task, if and when it is possible. At the level of the set. The dimension of consistency is developed across the data elements and could be user-defined as integrity [24]. The consistency may be analysed both at the level of expressions and at the level of the concepts that the expressions represent. The former relates directly to IR techniques (matching of strings), and the latter to the deeper structures of articles, which may be consistent even if the surface is inconsistent [61]. Consistency is necessary but not sufficient for correctness [2]. Representing with quality means representing consistently, and this refers to the extent to which definitions of elements are consistent among themselves. Consistency allows predictability, helping users realize what they will and will not get in a database search [18]. Consistency means uniformity and agreement in the processing of all information units. For this to be achieved, strict compliance with rules and working instructions is necessary [19]. Consistency favours quality, as can be proved by the fact that in an excellent database all records are similarly structured [62]. Consistency has a medium relative importance of 16%, being strongly linked to the RP-planning factors of synthesizing and structuring within the next deployment (Figure 3).

4.2.3. Accuracy. The IP-design dimension of accuracy has been strongly linked to the IP-planning factor of representing. As a consequence, a relative importance of 16% is deduced.

The basic entity model [63] conceives data as triple \((e, a, v)\), where \(e\) is an entity in a conceptual model, \(a\) is an attribute of entity \(e\), and \(v\) is a value from the domain of attribute \(a\). According to this model, the accuracy of a datum refers to the degree of closeness of its value \(v\) to some value \(v'\) in the attribute domain considered correct for the entity \(e\) and the attribute \(a\) [2]. Though accuracy may be expressed as a measure of inaccuracy, quantification of data inaccuracy is a non-trivial task, if and when it is possible. At the level of
the abstracting attribute, the factor accuracy refers to the extent to which the abstract correctly represents the original text. A theme covered in the abstract could be an inaccurate representation of the original because of an intellectual error or an error of carelessness [64]. Accuracy means the avoidance of errors at all stages of creating an information unit [19]. Different levels of accuracy may be needed in different circumstances. For example, audits of financial data, assessments of the extent of air pollution to use in environmental performance measures, and opinion surveys may all require different levels of accuracy [65]. Absence of errors includes such aspects of the source document as capitalization, non-English characters, character attributes, symbols, chemical formulae, and figures [54].

With a medium relative importance of 16%, accuracy will be strongly linked to the RP-planning factors of comprehension and synthesizing in the next deployment (see Figure 3).

4.2.4. Comprehensiveness. Looking back, the IP-design dimension of comprehensiveness has been strongly linked to the IP-planning factor of representing. Its deduced relative importance is 16%.

An excellent database covers its subject area comprehensively [21], including most or all aspects and consequently having a wide scope. Internally, the comprehensiveness dimension is directly linked to the relevance dimension. At the record level it measures the degree of breadth of the data elements that make up the record [17]. The dimension of comprehensiveness depends on the attributes to which it refers, being directly related to exhaustivity.

With a medium relative importance of 16%, comprehensiveness will be strongly linked to comprehension, moderately to selecting and weakly to synthesizing at the next stage of RP planning (Figure 3).

4.2.5. Format. The dimension of format has been strongly linked to the relating and moderately to the representing IP-planning factors. Consequently, its relative importance is 17%.

From a user perspective, the most important data representation issue concerns the means chosen to represent data values by symbols in a recording medium, called the format [66]. Formatting transforms the document model representation into an output model representation (the abstraction of the logical structure of the documents into an abstraction of their physical appearance [67]). Format is strongly related to accessibility, which extends to what potential additional access points are present in the database records [18].

With a medium relative importance (17%), the IP-design dimension of format will be strongly linked to structuring and weakly to the comprehension and synthesizing RP-planning factors (Figure 3).

4.2.6. Currentness. The currentness dimension at the stage of IP design is strongly linked to the seeking factor at the time of IP planning. In the sense of being within the right timeframe, currentness for a datum may be expressed as a measure of how outdated the datum’s attribute is [2]. Similarly, currency-timeliness is the period between the publication date and the appearance of the information unit of this publication in a database. Compared to the other dimensions, the relative importance of the currentness dimension is low at 6% (see Figure 2). Looking forward, the currentness dimension in IP design will be strongly linked to the selecting RP-planning factor (Figure 3).

4.3. RP-planning factors

Once we have developed the factors and dimensions of IPs, the third deployment directs the research toward the definition of the RP-planning factors. During this third stage of the S-QFD the six IP-design dimensions (WHAT) are translated into four RP-planning factors (HOW): comprehension, synthesizing, structuring, and selecting. The strength of the relations among the WHATs and the HOWs is shown in Figure 3.

4.3.1. Comprehension. The RP-planning factor of comprehension is strongly linked to relevance, accuracy, and comprehensiveness, and weakly to format at the stage of IP design. As a consequence of these relationships, its derived relative importance (32%) is significantly high.

Comprehension can be broken down into three stages: perception, which concerns translation from sound to a word representation; parsing, that concerns translation from the word representation to a meaning representation; and utilization, which concerns the use to which the comprehension puts the meaning of the message. Comprehension of a text depends critically on the perceiver’s ability to identify the hierarchical and causal structures that organize it. The output of the comprehension process is represented as a set of propositions that encode the information and the text and the inferences made from the text. A major limitation of comprehension is that only a few propositions can be held active in the working memory at any given
time. New theories define textual comprehension as the process of creating a mental model that serves to interpret the facts described. In order for this to happen, there is a need for the memory participation to implement inferences based on the previous knowledge of receivers and the strategic-contextual factors. Cognitive models gather the way in which humans understand, organize and represent information, taking into account their knowledge of discursive structures and mental representations [41].

With a high relative importance of 32%, the factor of comprehension is the most important at the RP-planning stage. As can be seen, it is strongly linked to the human resources dimension, and moderately to software within RP design (Figure 4).

4.3.2. Synthesizing. The RP-planning factor of synthesizing has been strongly linked to the IP-design dimensions of consistency and accuracy, moderately to relevance, and weakly to comprehensiveness and format. As far as its relative importance is concerned (25%), it occupies the second place among the RP-planning factors.

Synthesizing means the combination of separate elements of thought into a whole, in other words, the move from simple conceptions to complex ones. Analysis and synthesis, though commonly treated as two different methods, are, if properly understood, simply the two necessary parts of the same method: each is relative and dependent on the other. Though the processes of synthesis have to be entropic, coherent and balanced, it is practically impossible to establish techniques of synthesis that are valid for all types of documents [68].

Its relative importance within the RP-planning factors is 25%, being functionally linked to the RP-design dimensions: strongly to human resources and moderately to software (Figure 4).

4.3.3. Structuring. The RP-planning factor of structuring is strongly linked to the IP-design dimensions of consistency and format, and weakly to relevance. As a result, its relative importance among the RP-planning factors is 23%.

Structuralism is a sociological, anthropological and linguistic theory stressing that human actions are guided by beliefs and symbolic concepts, and that underlying these are structures of thought which find expression in various forms. The object of study is therefore to uncover the structures of thought and their influence in shaping the ideas in the minds of the human actors who created the record. From the pragmatic perspective of this research, structuring is a two-sided processing body: on the cognitive side, structuring means a complex composition of knowledge as elements and their combinations; on the physical side, structuring involves the manner of construction of something and the arrangement of its parts. These two points of view, as two sides of the same coin, are especially linked in the case of data: structural shape has to be especially adapted to structural content. Within this double line of structuralism and functionalism – the scientific tradition that stresses the relationship between a physical structure and its function – a database developer has to be a person who forms structures, that is, a structurist.

With a relative importance of 23%, the RP-planning factor of structuring is strongly linked to the RP-design dimensions of software and weakly to human resources and tangibles (Figure 4).

4.3.4. Selecting. The RP-planning factor of selecting has been strongly linked to the relevance and currentness and moderately to the comprehensiveness IP-design dimensions.

Denoting a choice, selection is the process of
purposive elimination. Developed by means of contraction, reduction and condensation strategies, the aim of selection is to retain only that information which is relevant. From the huge amount of information contained in the original sources, the selection processes act by eliminating non-relevant information and retaining the relevant. Though several kinds of software may meet, to a limited degree, specific pre-determined criteria for selecting resources, only experienced subject experts possess the level of knowledge required to select high-quality resources [69].

The relative importance of selecting obtained through the deployment is 20%. Functionally speaking, the RP-planning factor of selecting is strongly linked to human resources, moderately to software, and weakly to tangibles (Figure 4).

4.4. RP-design dimensions
The fourth deployment of the S-QFD model provides the transformation of RP-planning factors into RP-design dimensions. The four RP-planning factors have been translated into three dimensions at the time of the RP-design: human resources, software and tangibles (Figure 4).

4.4.1. Human resources. The RP-design dimension referred to as human resources has been strongly linked to the RP-planning factors of synthesizing, comprehension, and selecting, and weakly to structuring. As a consequence of these relationships, it has high relative importance (56%).

The needed expertise of a human expert involves intangible information embodied in human memory, knowledge experience or skill [70]. This leads us to a consideration of a form of literacy appearing from the outset to be based on rather wider premises than one or more skills: information literacy [71]. The term information literacy, sometimes referred to as information competency, is generally defined as an understanding and set of abilities enabling individuals to ‘recognise when information is needed and have the capacity to locate, evaluate, and use effectively the needed information’ [72]. Being information literate requires knowing how to clearly define a subject or area of investigation; select the appropriate terminology that expresses the concept or subject under investigation; formulate a search strategy that takes into consideration different sources of information and the variable ways that information is organized; analyse the data collected for value, relevancy, quality, and suitability; and subsequently turn information into knowledge [73]. A report of the US National Research Council [74] promotes the concept of ‘fluency’ with information technology. While information literacy focuses on content, communication, analysis, information searching and evaluation, information technology ‘fluency’ focuses on a deep understanding of technology and graduated – increasingly skilled – use.

The relative importance obtained for the human resources dimension (56%) is significantly high and deserves special attention. At the RP-design stage, people are the main dimension.

4.4.2. Software. The RP-design dimension of software is strongly linked to the structuring, and moderately to the synthesizing, comprehension, and selecting RP-planning factors. Regarding these relationships, its derived relative importance is 40%.

Software is a generic term for organized collections of computer data and instructions, often broken into two major categories: while system software provides the basic non-task-specific functions of the computer, controlling, integrating and managing the individual hardware, application software is used to accomplish specific tasks other than just running the computer system. Software includes both source code written by
humans and executable machine code produced by assemblers or compilers. Situated at the frontier between people and machines, software is their intermediary, providing the unavoidable human–computer interaction. One of the problems of software is that in the endless dialogue between the designer and the user of a specific human–computer interaction, one of the two participants is almost irremediably virtual when the other is present [44].

Considering software as a prolongation of human capabilities, it is not surprising that its relative importance derived from our S-QFD should be considerably high (40%).

4.4.3. Tangibles. Tangibles include the physical evidence [75]. The RP-design dimension of tangibles involves all the physical facilities and the equipment implied in the design of the processes of raw data representation. Original documents, spaces, equipment and, above all, the hardware necessary to carry out the design and production of the data are the main ingredients of the tangibles dimension, making possible the bridge between the virtual and physical worlds. The RP-design dimension of tangibles is weakly linked to the structuring and selecting RP-planning factors. Its relative importance is significantly low (4%).

5. Conclusions

We consider that the main contribution proposed by this paper is an unprecedented picture of the key factors and dimensions driving the data representation topic from the unprecedented QFD perspective. This research will clarify other classic points of view related to the factors involved in the planning and the dimensions related to the design of data representations.

(1) At the time of IP planning, DBs have to be truly representative of the original sources. The prevailing factor is based on the representing function, whose relative importance rises to 54%. With regard to their relative value, the other three deployed factors are less worthy – relating (18), filtering (16), and seeking (12) – with a similar relative assessment. But the representing and the filtering IP-planning factors correlate in a negative way (see Figure 1).

(2) When designing IPs, six dimensions have been deployed. Relevance (30%) is the core or central dimension around which the others are swinging, not only for the databases as a whole but also regarding individually their data elements.

Format (17), comprehensiveness (16), consistency (16), and accuracy (16) are also fundamental dimensions to be taken into account in the IP design. Although with a lower rank, currentness (6) is to be considered as well (Figure 2).

(3) Within the RP-planning stage, database planners have to place special emphasis on the comprehension factor (32%). However, the relative importance of the other three factors derived from our deployment, synthesizing (25), structuring (23), and selecting (20), must not be forgotten (Figure 3).

(4) Human resources (56%) remain the most important dimension when designing RPs. Considering the interaction between human beings and software, the large relative importance of this (40) is not surprising. Compared to the other two RP-design dimensions, tangibles (4) prove to be almost irrelevant (Figure 4).

(5) This paper is conceived with the aim of offering a new starting point in the data representation issue, and a significant amount of future research is required before this framework can become robust. The field of data representation will be better structured if those involved keep in mind the four S-QFD stages of this model. Beyond the detection and location of these functional forces (factors-dimensions) driving the data representation there is work to do whose nature is unpredictable.

Taking into account that our purpose has been limited to the conception – planning/design – of the main factors and dimensions involved in the raw data representation issue, one has to stress that this is not the place to find knowledge other than the discovering, framing and structuring of these factors and dimensions, that is to say, the modelling of a data representation quality framework. This framework is limited to the concepts introduced here. As we are still at a reflective stage of the research, it is very early to assess how these conceptual and theoretical results can be used practically. Data representation planning and design still remain more of an art than a science, as it takes a large amount of creativity and vision to design a solution which is robust and usable and can stand the test of time [76]. Expert systems can never truly emulate the reasoning process of a human expert, nor can they learn from previous design sessions as a human expert would [77]. The reliability and quality of databases and their data still depend on people and their choices. Thus, human resources are by far the most important data condition.
Acknowledgement

This research was financed by the Junta de Andalucía, Consejería de Educación y Ciencia, 2003, within the Ciberabstracts project.

References

[32] P. Hernon and J.R. Whitman, Delivering Satisfaction and


