Measurement of the Quality of an FCO-IM Conceptual Data Model

Fazat Nur Azizah[1], Guido P. Bakema[2], Benhard Sitohang[1], Oerip S. Santoso[1]

[1]School of Electrical Engineering and Informatics, Bandung Institute of Technology
Jln. Ganesha no. 10, Bandung, Indonesia
[2]Faculty of Engineering, Institute of Information Technology, Media and Communication,
HAN University of Applied Sciences
Ruitenberglaan 26, 6802 CE, Arnhem, The Netherlands
fazat@stei.itb.ac.id
benhard@stei.itb.ac.id
oerip@stei.itb.ac.id
guido.bakema@han.nl

Abstract: Data models provide the foundation to organization's activities since they support the organization's systems and data. Therefore, the quality of the data models is foremost. We describe a methodology to measure the quality of conceptual data models created using a fact oriented data modeling called Fully Communication Oriented Information Modeling (FCO-IM). The measurement method is based on the framework to measure the quality of conceptual model by Lindland et al. Four components are to be considered in the measurement: domain, model, language, and audience interpretation. The quality are measured on three aspects: syntactic quality (measured by syntax correctness), semantic quality (measured by feasible validity and feasible completeness), and pragmatic quality (measured by feasible comprehension). The method is then used to determine the quality of several FCO-IM conceptual data models that were created using a pattern language of conceptual data models, a new method in data modeling that we are currently researching. The method contributes in data modeling area by providing a quantitative and instructive way of measuring the quality of conceptual data models, especially in FCO-IM.

Keywords: conceptual model, conceptual data model, data modeling, FCO-IM, measurement, pattern, quality

1. Introduction

Information is one of the critical assets of a modern organization. Information is extracted from data stored in database systems. An aspect of data management is the definition of the structures of data. The structures of data are designed in an activity called data modeling and the results are data models. Data models provide the foundation to organization's activities since they support the organization's systems and data [20]. Therefore, the quality of the data models is foremost.

Data model is a collection of conceptual tools for describing data, data relationships, data semantics, and consistency constraints [16]. Three levels of data models are defined [19]: conceptual, logical, and physical data model. Conceptual data model is a relatively technology-independent specification of data structures and is close to business requirements [19]. We focus on conceptual data model rather than the logical or physical data model because a conceptual data model can be viewed as the translation of business requirements into technical form of the structures of data (thus, it serves a "link" between human and machine) and providing logical and physical data model is a matter of transforming the conceptual data model using established algorithms. Thus, the challenge is how to provide high quality conceptual data models.

Received: September 22, 2010. Accepted: December 20, 2010

The quality of conceptual data model is subject to ongoing discussions. There is little agreement on what constitutes a high quality conceptual data model. Nevertheless, a lot of authors provide a list of properties of the so called high quality conceptual data model (see for example: [12],[15],[18]-[20]). Unfortunately, only a few of them actually provide a way to measure the properties. Thus, the quality of conceptual data model is most of the time subjectively measured, depending upon the "taste" of the people who do the measurement. A work by Husain et al. [12] provides a way to objectively measure the quality of an entity-relationship (ER) conceptual data model (for ER model see [16]), but it is only on a property called completeness and it does not provide a whole discussion on the aspects of the quality of conceptual data model.

The objective of this paper is to describe a methodology to measure the quality of a conceptual data model, especially FCO-IM conceptual data model. FCO-IM (*Fully Communication Oriented Information Modeling*) [3],[7] is a conceptual data modeling method which belongs to the fact oriented data modeling (FOM) approach. The use of FCO-IM in this work is based on two reasons:

1. As a conceptual data modeling method, FCO-IM is known for its preciseness in modeling a universe of discourse (UoD) because it is based exactly on how users communicate the UoD [7] in comparison to other conceptual data modeling methods, although it is less popular in comparison to ER modeling or object-oriented modeling [8] or even ORM [9], another FOM method. It is even equipped with a way to regenerate the expressions used by the users to communicate the UoD based on the conceptual data models (see section 2.1), which provides a way for the domain experts to validate a model. We will use these advantages to measure the quality of the conceptual data models.
2. We are conducting a research on pattern language of conceptual data model patterns based on FCO-IM. The use of pattern language of conceptual data model patterns in a conceptual data modeling activity is aimed at providing high quality conceptual data models [5],[6]. Thus, we need a way to measure the quality of the resulting conceptual data models that are modeled using the pattern language.

The measurement of the quality of conceptual data model is based on the work by Lindland et al. [14]. Lindland et al. provides a clear-cut framework to measure the quality of conceptual model in general and it is based on an extensive study on all aspects of the quality of conceptual models. Thus, it works also for conceptual data model. The problem is: since it is a generic framework, it does not provide the details required to measure the quality properties.
This paper contributes in providing a quantitative method to measure the quality of FCO-IM conceptual data model based on the general framework for measuring the quality of conceptual models proposed by Lindland et al. We provide the details required in Lindland et al.'s framework to measure the quality of an FCO-IM conceptual data model. Using the method, the measurement of the quality of a conceptual data model can be carried out objectively and thus, the results become more reliable.

2. Foundations and Related Works

A. FCO-IM

Fully Communication Oriented Information Modeling (FCO-IM) is a *fact oriented modeling* method created based on NIAM (Nijssen's Information Analysis Method) and can be considered as an extended NIAM. In FCO-IM, information analysis is carried out on fact expressions, i.e. sentences that express how users communicate concrete facts of a Universe of Discourse (UoD). The final product of data modeling using FCO-IM is called an *Information Grammar (IG)*, which is considered as the conceptual data model. An IG stores the fact expressions in type level. These are called the fact types. Fact types are accompanied by data model constraints which are basically the rules that define valid fact expressions. Parts of a fact type are called roles. Roles of a fact type can be played by either an object type

(a representation of real world object) or a label type (a representation of a group of values). Object type is considered as a nominalized fact type. To help user to understand an IG better, an *Information Grammar Diagram (IGD)* is used. Further description of FCO-IM can be found in [3],[7].

Consider the following examples of fact expressions:

The name of product PAP192 is Johnson paper.
" " " " PEN202 " Goldstein pen.
" " " " DSK401 " Jerry's disk.

Suppose there are some rules that work on the facts as the following:

Product is uniquely identified by product code.
Every Product must have a name.
Every Product is assigned only one name.

An IG can be considered as an abstraction of concrete fact expressions. The abstraction is carried out by taking into account only the common parts of fact expressions to form a fact type. For example: from the fact expressions, we can create the following IG:

Name of Product
F2 : "The name of <Product : O2> is <product name>."
O2 : 'product <product code>'
UC3 : "Name of Product is uniquely identified by Product."
UC4 : "Product is uniquely identified by product code."
TC2 : "Every Product must be present in Name of Product."

The IG consists of the followings:
- F2 is a fact type called Name of Product.
- F2 has two roles (the parts between < >). The first role is played by object type Product (which is expressed using object type expression O2). The second role is played by a label type product name.
- UC3 and UC4 are constraints involved in the IG. Both are called uniqueness constraints. A uniqueness constraint defines that the values that may be filled in particular role(s) must be unique. UC3 ensures that every Product is assigned only one name, while UC4 ensures that every Product is uniquely defined by product code.
- TC2 is a totality constraint which states that every Product must have a name. This constraint ensures that every Product must have a product name.

The IGD for the example is shown in Figure 1. Roles are presented as rectangles with unique number in them. The roles conceive fact types, for instance: F2 consists of role #5 and #6. Object type, in this case: Product, is shown by a circle surrounding some roles. Label types,

in this case: product name, are shown as dash-lined circles. Uniqueness constraints UC3 and UC3 are presented as two-way arrows over roles. Totality constraint TC2 is represented as a dot in the Product end of the connecting line between Product and role #5.

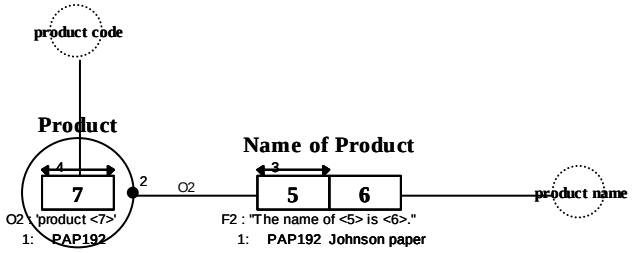

Figure 1. An example of IGD

A proper IG should be able to be used to regenerate the fact expressions from which the modeling started. In this manner, an IG can be validated against the facts given by the domain experts. For example: suppose we provide the value PAP192 for product code and Johnson paper for product name we will have the following fact expression regenerated:

The name of product PAP192 is Johnson paper.

The capability to regenerate the fact expressions based on the conceptual model is used in this paper to provide the details to measure the quality of an FCO-IM conceptual data model (IG). An IG must satisfy the IG well-formedness rules (see [7]) which work as the syntax guideline in creating an IG. Some of them are:

1. Each label type or fact type, nominalized or not, must be given a name which is unique throughout the IG.
2. Each role must have a unique designation (for example: a role number).
3. Each role must be part of a fact type.
4. Each role must be played by exactly one label type or nominalized fact type.
5. Each label type and each nominalized fact type must play at least one role.
6. Each fact type that is not nominalized, must have at least one fact type expression, in which the role designations (see rule 2) recur. These role designations replace the label type names and the object type names with object type expression unique designations (see rule 9).
7. An existence postulating fact type expression may be given to a nominalized fact type, but this is not always required.
8. Each nominalized fact type must have at least one object type expression, in which the role designations (see rule 2) recur.
9. Each fact type expression is given a unique designation (for example: F1, F2, and so on), and each object type expression gets a unique designation (for example: O1, O2, and so on).
10. A role must be connected to a nominalized fact type via at least one object type expression.

B. Quality of a Conceptual Model

The quality of an FCO-IM conceptual data model is measured based on a framework on quality of a conceptual model by Lindland et al. [14]. According to the framework, there are four essential elements of a conceptual model (language, domain, model, and audience

interpretation) and three connecting aspects (syntax, semantic, and pragmatic), as shown in Figure 2.

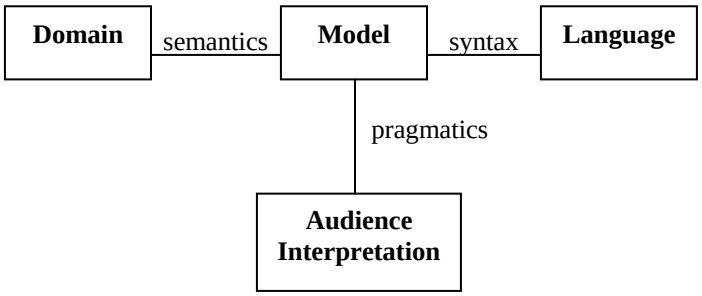

Figure 2. Elements and connecting aspects of conceptual model quality framework by Lindland et al. [9]

The elements are described as the following [14]:

1. Language (L)
   Language (L) consists of all statements that can be made according to the syntax, which for most languages is an infinite number. The alphabet contains a set of modeling constructs, while the grammar contains rules to define how to legally combine the modeling constructs.
2. Domain (D)
   Domain (D) consists of all possible statements that are correct and relevant for solving the problem. Domain is the ideal knowledge of a particular problem.
3. Model (M)
   Model (M) is a set of statements that is actually created. Two components of M are:
   - Explicit model: the set of statements explicitly made.
   - Implicit model: the set of statements that can be derived from explicit model according to L's deduction rules. This set usually contains infinite number of statements which are generally redundant.
4. Audience Interpretation (A)
   Audience interpretation (A) is a set of statements on what the audience thinks the model M contains. Audience is every party who needs to understand the model including all stakeholders in the development process (can be customers, domain experts, analysts, designers, even computers which must understand the model to automatically manipulate it).
5. Syntax
   Syntax relates the model (M) to language (L) by describing relations among language constructs without considering their meaning. Syntax is used to check whether a model is according to the constructs and rules of the language.
6. Semantics
   Semantics relates the model (M) to domain (D) by considering, not only syntax, but also relations among statements and their meaning. Semantics is used to check whether the model is according to the domain to be modelled.
7. Pragmatics
   Pragmatics relates the model (M) to audience participation (A) by considering, not only syntax and semantics, but also how the audience interprets the model. Pragmatics is used to check whether the model is well comprehended by the audience.

The framework proposes three quality aspects to be measured [14]:

1. Syntactic quality
   The goal of syntactic quality is the syntactic correctness. This means that all statements in M are according to the syntax in L, that is $M\backslash L = \varnothing$.
2. Semantic quality
   The goals of semantic quality are validity and completeness. Validity means that all statements made by the model M about the domain D are correct and relevant, that is $M\backslash D = \varnothing$. Completeness means that the model M contains all statements about the domain D that are correct and relevant or $D\backslash M = \varnothing$. Nevertheless, it is almost impossible achieve total validity and completeness. Thus, the framework introduces the concept of feasible validity and feasible completeness.
   Feasible validity is defined as $M\backslash D = E \neq \varnothing$, in which there is no e in E (E is a set of invalid statements in the model) such that the additional benefit to the model M from removing e is greater than the drawbacks of eliminating the invalidity.
   Feasible completeness is defined as $D\backslash M = F \neq \varnothing$ in which there is no statement f in F (F is the set of correct and relevant statements not yet in the model M) such that the additional benefit to the model M from including f exceeds the drawbacks of including it.
3. Pragmatic quality
   The goal of pragmatic quality is audience comprehension. For this purpose, the framework introduces the concept of model projections M_1, M_2, …, M_n and a corresponding set of audience projections A_1, A_2, …, A_m. The projections represent different group's interpretations. Comprehension means that all model projection have been understood by their relevant audience, that is $(\forall_i)$ $(M_i = A_i)$. It means that all audience members completely understand the statements in the model that are relevant to them.
   The framework argues that it is not realistic, especially in large models, to expect that every member of audience to comprehend everything in the model completely. Thus, it introduces the concept of feasible comprehension which is defined as $(\exists_i)$ $(A_i\backslash M_i) \cup (M_i\backslash A_i) = G \neq \varnothing$. It means there is no statement g in G such that the benefit of rooting out the misunderstanding corresponding to g exceeds the drawbacks of taking that effort (G is the set of statements in the model relevant to a particular audience group that has been misunderstood by the group plus the statements the audience believes in the model but is not).
   Lindland et al. argue that although the framework seems to simplify a lot of criteria of the quality conceptual model, it covers all of them [14]. For instance: correctness is covered by validity, consistency is covered by validity and completeness, etc.

C. Related Works

A lot of authors present their ideas about what constitutes a high quality conceptual data model. For instance: Simsion et al. provides 10 properties of a good conceptual data model, such as: scope/coverage, non-redundancy, enforcement of business rules, etc. [18]-[19]. West gives 7 criteria of a high quality data model, such as: meet data requirement, be clear and unambiguous, etc. [20]. Reingruber et al. provides another list of the properties of high quality data model [15]. None of these authors present exactly how to measure each of the properties.
Husain et al. described a way of measuring the completeness of an ER conceptual data model using effort-based approach [12]. Nevertheless, the work gives only how to measure completeness and does not discuss all aspects of the quality of conceptual data model.

3. Measuring the Quality of an FCO-IM Conceptual Data Model

It should be clear by now that there is a need to provide a way to measure quantitatively a conceptual data model. In this paper, we focus on FCO-IM conceptual data model and the measurement is based on the work by Lindland et al.

The framework presented by Lindland et al. discusses all aspects of the quality of conceptual model based on an extensive study on other works on quality conceptual models. They compared various quality properties described by different authors and concluded that there are 3 important aspects to measure the quality of a conceptual model: the syntactic quality, the semantic quality, and the pragmatic quality (see further discussions). In our opinion, this work provides a concise yet powerful means to measure the quality of a conceptual model that can be extended to conceptual *data* model. Nevertheless, further details on how to use the framework to measure a conceptual data model is required to be investigated. We combine this with FCO-IM. In FCO-IM, currently there is no work that addresses formally the quality of the resulting conceptual data model.

A. Elements of Measurement Model

We extend the framework of conceptual model proposed by Lindland et al. to measure the quality of an FCO-IM conceptual data model, or the Information Grammar (IG). We define the elements of the framework as the following:

1. Language (L)
   L consists of all statements that can be made according to the syntax of FCO-IM. The alphabet of the FCO-IM consists of the constructs of FCO-IM, such as fact type, nominalized fact type (object type), label type, etc. The grammar is in general described by IG well-formedness rules (see section 2 and further in [7]).
   The set of the statements in IG well-formedness rules is denoted with L and each statement is denoted as l_i with i = 1, 2, ..., n. Thus: L = { l_1, l_2, ..., l_n }.
2. Domain (D)
   D contains all possible statements that are correct and relevant to the modelled UoD. It is a set of elementary fact expressions as well as business rule statements on the UoD that are given by the domain experts. The example of such statements can be found in section 2.
   The set of the statements in a domain is denoted with D and each statement is denoted as d_i with i = 1, 2, ..., n. Thus: D = { d_1, d_2, ..., d_n }.
3. Model (M)
   M consists of the statements of an FCO-IM conceptual data model that is used to model the UoD. Two components of M are:
   - Explicit model (M_E): a set of statements stated in an IG.
   - Implicit model (M_I): a set of statements that can be created based on the IG.

In our research, we define M_E in two forms which are equal ($M_{E1} \equiv M_{E2}$) :

- M_{E1}: the IG itself.
  Each statement in M_{E1} is denoted as m_{E1i} with i = 1, 2, ..., n. Thus: M_{E1} = { m_{E11}, m_{E12}, ..., m_{E1n}}.
- M_{E2}: the sentences and business rules regenerated from the IG.
  Each statement in M_{E2} is denoted as m_{E2j} with j = 1, 2, ..., m. Thus: M_{E2} = { m_{E21}, m_{E22}, ..., m_{E2m}}.

M_I consists of other sentences and business rule statements that can be derived from the IG. We define M_I also in two forms which are equal ($M_{I1} \equiv M_{I2}$):

- M_{I1}: derivable fact types and constraints.
  Each statement in M_{I1} is denoted as m_{I1k} with k = 1, 2, ..., p. Thus: M_{I1} = { m_{I11}, m_{I12}, ..., m_{I1p}}.
- M_{I2}: the sentences and business rule statements that can be generated based on the fact types and constraints in M_{I1}.
  Each statement in M_{I2} is denoted as m_{I2l} with l = 1, 2, ..., q. Thus: M_{I2} = { m_{I21}, m_{I22}, ..., m_{I2q}}.

Thus, explicit or implicit model comes in two types:
- M_{X1}: the actual model, stated in FCO-IM conceptual model syntax.
- M_{X2}: the sentences and business rule statements that can be regenerated based on M_{X1}.

4. Audience interpretation (A)

Based on the type of model (M_{X1} and M_{X2}), we divided the audience into two groups:
- *Domain experts*, i.e. the audience that concerns mainly to the meaning of the model (M) without having to consider the notations used in the model. Their interests are to know whether they can relate the knowledge they know about a UoD with the knowledge they can get from the model (M).
  The model relevant for this group of audience is M_{E2} and M_{I2}, since they are in the form of sentences that state the facts and business rule statements which are appropriate to understand the meaning of the model. If the model is projected into several projections, there will be several subgroups of domain experts; each corresponds to one model projection.
- *Information analysts*, i.e. the audience that concerns not only to the meaning of the model (M), but also the syntax of the model (M). They can relate the model not only to the UoD, but also to the modeling language (L).
- The model relevant for this group is M_{E1} and M_{I1}. If the model is projected into several projections, there will be several subgroups of information analysts; each corresponds to one model projection.

The set consisting of the audience interpretations is denoted as A. Suppose there is a model projection M_i with i = 1, 2, ..., n, then A_i is the audience comprehension of M_i. Thus: A = { A_1, A_2, ..., A_n }.

5. Syntax
   Syntax connects an IG (FCO-IM conceptual data model) with the constructs and grammar of FCO-IM. The rules on FCO-IM constructs and grammar are summarized into IG well-formedness rules (see again section 2.1).
6. Semantics
   Semantics relates the IG with the statements on facts and business rules of a UoD.
7. Pragmatics
   Pragmatics relates the IG with audience interpretation from each group.

B. Measuring the Quality Aspects

To measure the quality aspects (syntactic, semantic, and pragmatic quality) of an FCO-IM conceptual data model, each quality aspect, described in section 2.2 and further in [14], is adapted as the following:

1. Syntactic Quality
   The goal of syntactic quality is the syntactic correctness. In our research, we check the syntactic correctness of M using in particular the IG well-formedness rules (see section 2.1 for several examples and further in [7]). To check the syntax of a model, we consider only M_{E1} and M_{I1}. The statements of the IG well-formedness rules that are obeyed by a model M is denoted as L_M. The syntax correctness (SyC) is calculated as the ratio between the number of statements in L_M and the number of statements in L, presented in percentage, as the following formula:

$$SyC = \frac{(L_M)}{|L|} * 100 \quad (1)$$

2. Semantic Quality
   The goals of semantic quality are feasible validity and feasible completeness. In our research, the validity is checked for every statement in M_{E2} and M_{I2} by finding in D a statement which is the same or semantically the same as the statement in M_{E2} (denoted as M_{E2D}) or in M_{I2} (denoted as M_{I2D}) . Thus, the feasible validity (SeFV) is measured as ratio between the number of statements in M_{E2D} and M_{I2D} and the number of statements in M_{E2} and M_{I2}, presented in percentage, as the following formula:

$$\text{SeFV} = \frac{|M_{E2D}| + |M_{I2D}|}{|M_{E2}| + |M_{I2}|} * 100 \quad (2)$$

   The completeness will be checked for every statement in D by finding in M_{E2} or M_{I2} a statement which is the same or semantically the same as the statement in D (denoted as D_M). The feasible completeness (SeFC) is measured as the ratio between the number of statements in D_M and the number of statements in D, presented in percentage, as the following formula:

$$SeFC = \frac{|D_M|}{|D|} * 100 \quad (3)$$

3. Pragmatic Quality
   The goal of pragmatic quality is feasible comprehension. The audience comprehension of a subject x for model projection M_i is denoted as PFC_{ix}. It is defined as the ratio between the number of the statements in M_i that are correctly comprehended (M_{ix}) and the number of the fact stated sentences and business rule statements in the model projection M_i, presented in percentage, as the following formula:

$$PFC_{ix} = \frac{|M_{ix}|}{|M_i|} * 100 \quad (4)$$

   The total feasible comprehension of model projection M_i, denoted as PFC_i, is the average of the audience feasible comprehension of m subjects is then:

$$PFC_i = \frac{\sum_{x=1}^{m} PFC_{ix}}{m} \quad (5)$$

   The feasible comprehension of a model M, denoted is PFC, is thus the average of PFC_i of n model projections:

$$PFC = \frac{\sum_{i=1}^{n} PFC_i}{n} \quad (6)$$

4. A Pattern Language of Data Model Patterns based on FCO-IM

The measurement methodology on the quality of conceptual data model is used to measure the quality of the data models resulting from the use of *a pattern language of data model patterns* based on FCO-IM. A data model pattern is defined as a relation between context,

problem, and solution; each of which concerns data modeling [6] which is based on the definition of pattern by Alexander [1],[2]. Other works on data model patterns include [10],[11],[17].

A data model pattern can relate to one another; forming the so called pattern language of data model patterns. The relationships among data model patterns are stated explicitly in a part of solution of a data model pattern which is called the *Information Grammar for Pattern (IG_P)*. An IG_P works as a template to create an FCO-IM conceptual data model, i.e. the Information Grammar (IG).
An example of an IG_P is as the following:

```
(attribute-of-object):
(F1):"[(expression-1)]<(G1#(1))>[(expression-2)]<(attribute's-
name)|(G1#(2))>[(expression-3)]."||
(F2):"[(expression-4)]<(G1#(2))>[(expression-5)]<(attribute's-
name)|(G1#(1))>[(expression-6)]."
(UC1):"(attribute-of-object) is uniquely identified by (G1#(1))."
[(UC2):"(attribute-of-object) is uniquely identified by (attribute's-name)|(G1#(2))."]
[(TC1):"Every (G1#(1)) must be present in (attribute-of-object)."]
[(TC2):"Every (G1#(2)) must be present in (attribute-of-object)."]
```

The IG_P is a part of the solution of a data model pattern called *Attribute Pattern* which is a pattern that can be used when there are two objects in which one of the objects is an attribute/property of the other [6]. In this example, the parts denoted with the text G1 are the parts in which other patterns must be generated, in this case, patterns from category G1 [6] which are patterns that are aimed at modeling the identification of an object. The example shown in section 2.1 is created based on this IG_P.

A further discussion on data model patterns and the pattern language can be found in [6]. So far, we have developed 15 data model patterns. Some of them are briefly described in [5],[13]. The pattern language of conceptual data model patterns is aimed at providing high quality [FCO-IM] conceptual data models. The conceptual data model patterns are created based on context, problems, as well solutions of data modeling that are frequently found in majority of cases. Because of this, specific aspects of a data modeling case may not be covered by the data model patterns and must be handled specially by the data modelers.

The methodology described in this paper is used to test the results of the use of the pattern language of data model patterns in several case studies. Note that this methodology works in general for measuring the quality of FCO-IM conceptual data models, not only for the ones created using the pattern language.

5. Tests on the Quality of FCO-IM Conceptual Data Models

Using the concepts of the quality of FCO-IM conceptual data model described in section 3, we conducted tests on the quality of conceptual data models that were created using the pattern language of conceptual data model patterns. The tests were carried out over 14 case studies taken from [7].

A. Student's Project Case Study

One of the test cases is a case on student's project. Students of a particular school are given a list of projects from which they have to choose 3 preferences. In the end they are assigned to one of the projects based on their preferences.

A sample of fact expressions based on the case is as the following:

"There is a student Peter Johnson."
"The mentor of student Peter Johnson is BLC."
"The school is offering project P101."
"Project P101 is supervised by BLC."
"Project P101 concerns developing a timekeeping system."
"The first preference of student Peter Johnson is project P101."
"Student Peter Johnson was allocated project P101."

There are 16 business rule statements to consider. Some of them are as the following:

1 "A student is identified by his/her first name and surname."
2 "Each student has a teacher as personal student adviser (mentor)."
3 "Each student will be assigned to one project."
4 "Each student is assigned to one of the projects based on their preferences."

The fact expressions and business rule statements conceive the set D.

A model, say M, aka. an IG was created for the case study using the first version of the pattern language of conceptual data model patterns. Note that we consider only a model that can be created entirely only by using the pattern language. Thus, specific aspects of the case which cannot be provided by the use of the pattern language are not to be taken into account. Because of the specific requirement of the patterns, most of the fact expressions are altered. Nevertheless, we do not regard this as an important disadvantage, since the meaning of the sentences is fully retained. A part of the IG containing the fact types, which is a part of M_{E1}, is as the following:

Student:
F1:"There is a student <firstname> <surname>."
O1:'student <firstname> <surname>'
Mentor of Student:
F2:"The mentor of <Student:O1> is <Teacher:O2>."
Teacher:
O2:'<teacher id>'
Project:
F3:"There is a project <project code>."
O3:'project <project code>'
Supervisor of Project:
F4:"The supervisor of <Project:O3> is <Teacher:O2>."
Student's Preference:
O4:'preference <ordinal no> of <Student:O1>'
Description of Project:
F5:"The description of <Project:O3> is <description>."
Project of Student's Preference:
F6:"The project of <Student's Preference:O4> is <Project:O3>."
Allocation of Project:
F7:"The allocation for <Student:O1> is <Project:O3>."

The IGD is thus as shown in Figure 3.

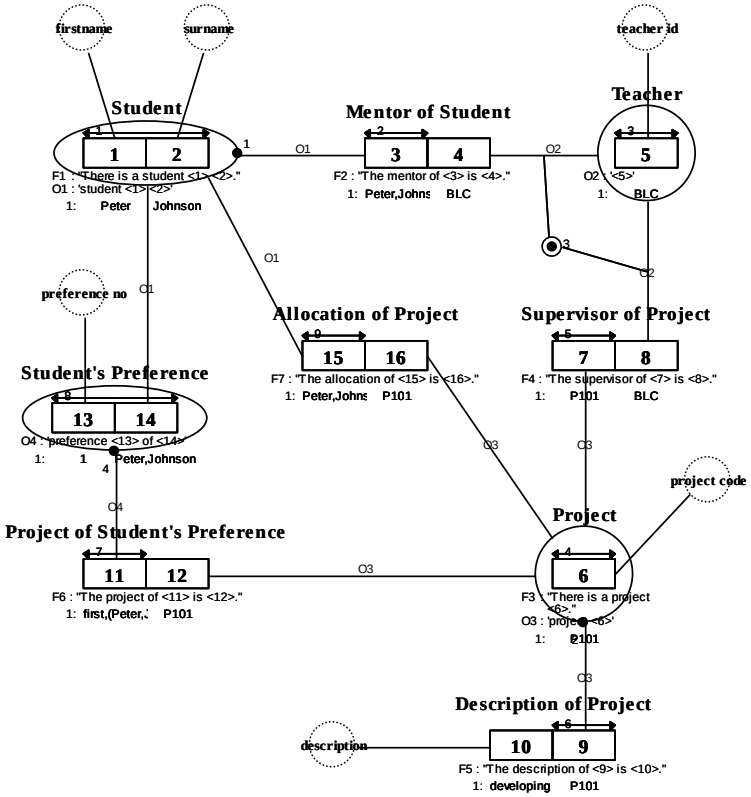

Figure 3. IGD for the student's project case

Based on the IG, the sample of fact expressions, which are a part of M_{E2}, are regenerated as the following:

"There is a student Peter Johnson."
"The mentor of student Peter Johnson is BLC."
"There is a project P101."
"The supervisor of project P101 is BLC."
"The description of project P101 is developing a timekeeping system."
"The project of preference 1 of student Peter Johnson is project P101."
"The allocation of student Peter Johnson is project P101."

Some constraints implemented within the model, also part of M_{E2}, are as the following:

a "Student is uniquely identified by firstname, surname."
b "Every Student must have Mentor of Student."
c "The Allocation of Student is uniquely identified by Student."

Constraint #a is a uniqueness constraint which states that every student is identified by his/her first name and surname, thus retaining rule #1. Constraint #b is a totality constraint which states that every student must have a mentor, thus, retaining rule #2. Constraint #c is another uniqueness constraint, which states that every student is allocated only to one project. This constraint states that each student will be assigned to one project, thus, keeping rule #3. There is no constraint in the model, however, which implements rule #4 which states that each student is assigned to a project based on his/her preferences. It actually requires the use of a subset constraint. This subset constraint is considered as a specific aspect of a case that none of the patterns supports. In the actual modelling process, this kind of detail must be handled directly by the modeller. But in this test, we want only to deal with everything that can be generated by the pattern language. The consequence is that there are some specific rules of the cases which are not covered by the models.

The quality of model M is calculated as the following:

1. Syntactic Quality
   There are 18 IG well-formedness rules to consider (10 of them can be observed in section 2.1). Thus, $|L| = 18$. Based on the observation to the model M, all 18 rules stand. Thus, $|L_M| = 18$. The syntactic correctness for the student's project is calculated as the following:

$$SyC = \frac{|L_M|}{|L|} * 100 = \frac{18}{18} * 100 = 100.00$$

   Thus, the syntactic correctness for the student's project case is 100.00%.
2. Semantic Quality
   To calculate the feasible validity and feasible completeness we consider the number of statements in D dan M. In the student's project, D consists of 7 groups of fact expressions and 16 statements of business rules, thus $|D| = 7+16 = 23$. For the student's case project, we consider only the explicit model (M_E), especially the M_{E2}. Thus $|M_{I2}| = 0$. There are 7 groups of fact expressions that can be regenerated from M_E and there are 13 statements of constraints that implement the business rules in M_E. Thus, $|M_{E2}| = 7+13 = 20$.
   For feasible validity in student's case project, all 13 statements in M_{E2} are relevant to statements in D, thus $|M_{E2D}| = 13$. Because $|M_{I2}| = 0$, $|M_{I2D}| = 0$. The feasible validity for the student's project is calculated as the following:

$$SeFV = \frac{|M_{E2D}| + |M_{I2D}|}{|M_{E2}| + |M_{I2}|} * 100 = \frac{13+0}{13+0} * 100 = 100.00$$

   Thus, the feasible validity of the student's project case is 100.00%.
   For the feasible completeness of the student's case project, all 7 fact expressions in D can be found in M. Nevertheless, from 16 business rule statements, only 13 of them can be found in M. Thus $|D_M| = 7+13 = 20$. The feasible completeness is calculated as the following:

$$SeFC = \frac{|D_M|}{|D|} * 100 = \frac{20}{23} * 100 = 86.96$$

   Thus, the feasible completeness for the student's project case is 86.96%.

3. Pragmatic Quality
   The model is provided to a group of 7 respondents with the domain expert qualification, say group A (see section 3.1). They are asked to compare the statements in M_E, especially M_{E2}, to the statements in D. A statement α in M_{E2} is said to be correctly comprehended by a respondent when the respondent can find the proper match between α with a statement β in D if such statement exists, or when he/she state that the statement cannot be found in D if such statement does not exist. For each respondent, the feasible comprehension is calculated based on the number of statements in M_{E2} that are correctly comprehended and the actual number of statements in M_{E2}. Afterwards, the feasible comprehension of the student's case project of respondents in group A is calculated from the average of the feasible comprehension of all respondents A, as the following:

$$PFC_A = \frac{\sum_{x=1}^{7} PFC_{Ax}}{7} = 96.83$$

   Because there is only one group of audience, PFC_A is also the feasible comprehension of the student's project case which is 96.83%.

B. Test Results

The syntactic quality and semantic quality were tested on all 14 case studies which are taken from [7], while the audience comprehension was tested only to 5 case studies and only to one group of respondents (as explained in section 5.1). The list of the case studies and the test results are presented in Table 1.

Table 1. Test results

Case Study	Test Results (in %)			
	Syntax qty.	Semantic qty.		Pragmatic qty.
	Syntax correctness	Feasible Validity	Feasible Completeness	Feasible Comprehension
Student's Project	100.00	100.00	86.96	96.83
Organization Chart	100.00	100.00	100.00	71.43
States	100.00	100.00	95.00	NA
Life Expectancy	100.00	100.00	66.67	NA
Furniture Emporium	100.00	100.00	87.50	NA
Health Care	100.00	100.00	89.29	NA
Education Institution	100.00	100.00	89.66	NA
International Wresting Competition	100.00	100.00	66.67	NA
Butter Company	100.00	100.00	96.43	NA
Employee	100.00	100.00	95.00	NA
Celestial Body	100.00	100.00	95.31	NA
Town Council at Oss	100.00	100.00	94.12	94.05
Boutique The Shirt Store	100.00	100.00	90.00	100.00
Speed Skating Champ.	100.00	100.00	96.97	100.00
Average	**100.00**	**100.00**	**88.04**	**92.46**

NA = not available

For syntax correctness, the average score is 100.00%. It means that so far, the resulting conceptual IGs are free from errors. For feasible validity, the average score is 100%. It means

that so far, all statements in the resulting conceptual data models are correct and relevant to the UoD.

For feasible completeness, the average score is 88.04%. It means that in average only 88.04% of the statements in the UoD can be modeled in the resulting IGs. Based on examination on the test results, the statements that cannot be modeled are specific business rule statements. All fact type expressions are modelled properly. As explained earlier, there are specific aspects of data modeling cases that cannot be contained in the data model patterns. Thus, the result is considered acceptable.

For feasible comprehension, the average score is 92.46%. It means that in average, the audience comprehended 92.46% of the model properly. Based on examination on the test results, all miscomprehensions of the model occur on the business rule statements. The probable cause is that the regenerated business rule statements, in the form of model constraint statements, are quiet different with the original statements of the business rules. Nevertheless, considering that all regenerated fact expressions are correctly comprehended by the respondents, it means that in general the models can be understood well.

Based on the test results, we can still consider that the resulting conceptual data models (i.e. the IGs) that were created using the pattern language of data model patterns are of high quality. There are two reasons to conclude this:

1. The numbers of the quality measurement parameters for all cases are above 50%. The average numbers for the tests are even beyond 75% with syntax correctness and feasible validity reaching 100%. We are not equipped with a limit value to state the boundary between high and low quality. We use 50% as the limit between high and low quality. An average number higher than 50% means that majority of the aspects in a data modeling case are covered.
2. The variation in the average numbers of feasible completeness and feasible comprehension is on the business rule statements which are translated into constraints in the model. In reality, modelers deal most of the time with the fact expressions rather than with the business rules. After all, the business rule statements are often not that specific. The case studies are taken from an academic book [7] in which a lot of details are taken into account. We believe that the pattern language of conceptual data model patterns can already be used to deal with real life data modeling cases.

Based on the test results, we go on further to develop our pattern language of conceptual data model patterns in order to improve the quality of the resulting conceptual data models especially in the feasible completeness and feasible comprehension aspects. The test will be run again to the results of the refinement.

Conclusions

According to Lindland et al. [14], the quality of a conceptual model can be measured on 3 quality aspects: syntactic quality, semantic quality, and pragmatic quality; based on 4 components: language, model, domain, and audience interpretation. The parameter of syntactic quality is the syntactic correctness; while the parameters of semantic quality are the feasible validity and feasible completeness; and the parameter of pragmatic quality is the feasible comprehension.

We use this concept to measure the quality of an FCO-IM conceptual data model (Information Grammar aka IG). We have defined ways to measure the syntactic correctness, feasible validity, feasible completeness, and feasible comprehension of an FCO-IM conceptual data model. The method was then used to measure the quality of FCO-IM conceptual data models which were created using the pattern language of conceptual data model patterns, a new method in data modeling that we are currently researching. The results of the tests lead to

a conclusion that the FCO-IM conceptual data models are of high quality, although remarks must be given to the values of feasible completeness and feasible comprehension.

The measurement method provides a quantitative way of measuring the quality of an FCO-IM conceptual data model. It makes it possible to determine whether an FCO-IM conceptual data model is of high or low quality in a more exact manner. This way of measurement is never been discussed anywhere before. This method also provides an instructive way of measuring the quality. We have defined what are the model, language, domain, as well as audience interpretation in the case of FCO-IM conceptual data model and how to measure the syntactic, semantic, and pragmatic quality. Users can just use the instruction to determine the quality of their FCO-IM conceptual data models.

However, the method is specific to FCO-IM conceptual data models. To measure the quality of conceptual data models which are created using ER or object-oriented modeling techniques, major changes are required to be undertaken. For other FOM methods, however, we expect that more little efforts are required since the nature of the modeling is similar.

The measurement method is expected to contribute in data modeling area, not only by providing a means to measure the quality of FCO-IM conceptual data models, but also a lesson-learned that the quality of a conceptual data model can be measured in an objective and more reliable manner. Further studies can be carried out especially on Lindland et al.'s framework to determine whether specific quality aspects of conceptual data models mentioned by several authors, such as: redundancy-freeness, elegancy, stability, flexibility, etc., are covered by this framework or not.

References

[1] Alexander, "The Timeless Way of Building", *Oxford University Press, USA*, 1979.

[2] Appleton, "Pattern and Software: Essential Concepts and Terminology", http://www.cmcrossroads.com/bradapp/docs/patterns-intro.html, accessed on 19/04/2006.

[3] "Articles on FCO-IM", http://www.casetalk.com/php/index.php?Articles, accessed on 18/3/2006.

[4] F. N. Azizah and G. Bakema, "Data Modeling Patterns using Fully Communication Oriented Information Modeling (FCO-IM)", *ORM Workshop 2006 (part of OnTheMove Federated Conferences and Workshops 2006), working papers, Montpellier, France*, 2006.

[5] F. N. Azizah, G. P. Bakema, B. Sitohang, and O. S. Santoso, "Generic Data Model Patterns using Fully Communication Oriented Information Modeling (FCO-IM)", *International Journal on Electrical Engineering and Informatics*, vol. 1, 2009.

[6] F. N. Azizah, G. P. Bakema, B. Sitohang, O. S. Santoso, "Information Grammar for Patterns (IG_P) for Pattern Language of Data Model Patterns Based on Fully Communication Oriented Information Modeling (FCO-IM)", *in press, 2010 ORM Workshop (part of 2010 OnTheMove Federated Conferences and Workshops), working papers, Crete, Greece*, 2010.

[7] G. Bakema, J. P. Zwart, and H. van der Lek, "Fully Communication Oriented Information Modeling (FCO-IM)", *HAN University, The Netherlands*, 2002. The book can be downloaded for free in http://www.casetalk.com/php/index.php?FCO-IM%20English%20Book.

[8] M. Blaha and W. Premerlani, "Object Oriented Modeling and Design for Database Application", *Prentice Hall*, 1998.

[9] T. Halpin and T. Morgan, "Information Modeling and Relational Databases", *Second Edition, Morgan Kaufmann, San Francisco*, 2008.

[10] C. Hay, "Data Model Patterns", *Dorset House Publishing, New York*, 1996.

[11] C. Hay, "Data Model Patterns: A Metadata Map", *Morgan Kaufmann Publishers, San Fransisco*, 2006.

[12] T. Hussain and M. M. Awais, "An Effort-based Approach to Measure Completeness of an Entity-Relationship Model", *Seventh IEEE/ACIS International Conference on Computer and Information Science*, 2008.
[13] Liem and F. N. Azizah, "Metadata Approach in Modeling Multi Structured Data Collection Using Object Oriented Concepts", *proceeding in International Conference on Networking and Information Technology (ICNIT)* 2010, June 2010, Manila, Philippines.
[14] O. I. Lindland, G. Sindre, and A. Solvberg, "Understanding Quality in Conceptual Modeling", *IEEE Software, March* 1994.
[15] M. C. Reingruber, W. W. Gregory, "The Data Modeling Handbook – A Best-Practice Approach to Building Quality Data Models", *John Wiley & Sons, Canada*, 1994.
[16] Silberschatz, H. F. Korth, and S. Sudarshan, "Database System Concepts", *Fourth Edition, McGraw Hill*, 2002.
[17] L. Silverston, "The Data Model Resource Book: Revised Edition", *Volume 1 dan 2, John Wiley & Sons Inc*., 2001.
[18] Simsion, "Better Data Models – Today: Understanding Data Model Quality", http://www.tdan.com/view-articles/5100, published 01/10/2005, accessed on: 5/11/2010.
[19] Simsion and G. Witt, "Data Modeling Essentials", *Third Edition, Morgan Kaufmann Publishers*, 2005.
[20] M. West, "Developing High Quality Data Models", *EPISTLE,* 1996, http://www.matthew-eest.org.uk/documents/ princ03.pdf, accessed on 04/07/2008.

Fazat Nur Azizah received a bachelor degree in informatics from Institut Teknologi Bandung (ITB), Indonesia in 2002 and a master degree in information systems development from HAN University of Applied Sciences, The Netherlands, in 2005. She is currently a student of doctoral program in School of Electrical Engineering and Informatics (SEEI), ITB, Indonesia. She is also a junior lecturer in SEEI, ITB. Her research interest is data modeling and metadata management. She can be reached at fazat@stei.itb.ac.id.

Guido P. Bakema is an emeritus professor in Data Architectures and Metadata Management in the Institute of Information Science, Media, and Communication in the Technical Faculty of HAN University of Applied Sciences, The Netherlands. He leads the research and competence group that focuses on proof-of-concept application of innovative approaches in industry, system houses, and governmental organizations. He will be available at guido.bakema@han.nl.

Benhard Sitohang is a professor in School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia. He obtained his undergraduate degree in electrical engineering at Institut Teknologi Bandung, his magister in Conception Assistee Des Systemes at USTL-Montpellier II, Montpellier, France, and doctoral degree in Automatique at Universite de Sciences et Techniques et Languedoc-Montpellier II, Montpellier, France. His field of interest is database systems. He will be available at benhard@stei.itb.ac.id.

Oerip S. Santoso obtained his medical doctor degree at Universitas Indonesia, Jakarta, Indonesia, his magister in computer science at University Wisconsin, Madison, USA, and doctoral degree in informatique at Universite Pierre et Marie Curie, Paris, France. He is currently an assistant professor in School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia. His field of interest is medical informatics. He will be available at oerip@stei.itb.ac.id.