Data Model Legibility
A comparison of 3 graphic styles

J.C.Nordbotten

Dept. of Information Science,
University of Bergen, N-5020 Bergen, Norway
e-mail Joan@ifi.uib.no

M.E.Crosby

Dept. of Information and Computer Science,
University of Hawaii at Manoa, Honolulu, HI 96822, USA
e-mail crosby@ics.uhics.hawaii.edu

ABSTRACT
Graphic models are used in information system analysis and design as communication tools between a systems analyst, who constructs the model, and a system user, who confirms that the model correctly depicts important requirements for the system. In addition to information types, structure, and behaviour, system models commonly describe integrity rules, or constraints, for the system. If the user is to be able to confirm the correctness of the model, it must facilitate understanding as well as attention to detail. Unfortunately, little is known about how users read and interpret graphic models, what level of detail is seen, or how graphic style influences perception.

In the following, we report from a pilot study in graphic data model perception. The models chosen for the experiment represent three different *graphic styles:* 1) fully graphic, 2) embedded structures, and 3) list based. Our observations indicate that graphic style influences the amount of detail seen and comprehended by the reader.

KEYWORDS graphic model perception, user interface, data modelling.

1. GRAPHIC MODELS as COMMUNICATION TOOLS

Graphic models of system specifications are used extensively as communication tools between users and designers of information systems to illustrate *information structure, behaviour, and constraint requirements* for an information system. The assumption is that the end-user will be able to read and understand the model and confirm that it correctly depicts important requirements for the system. However, little is known about how graphic models are read or interpreted, what level of detail is seen, or how particular graphic styles and symbols influence user perception.

Several research projects have studied the effectiveness of graphic representations of processes on programmer interpretation of system specifications under the hypothesis that presentation formats effect comprehension [Wrig73, Broo80, Petr95 among others]. An experiment which presented programmers with design specifications in three spatial formats: sequential, hierarchical, and branching flow charts, indicates that the combined use of a succinct program design language and a branching spatial arrangement significantly facilitated design interpretation [Shep82]. While a comparison of textual vs graphical presentations of 3-level nested program structures concluded that graphic presentations were not necessarily more accessible, comprehensible, or memorable than textual presentations. Particularly, novice graph readers suffered from misreadings and an inability to find and focus on essential parts of the graph [Petr95].

A recent study of the impact of graphic style on data model interpretation found that the graphic style of the 2 model types studied, NIAM [Nijs89] and EER [Tore86], had no significant impact on user comprehension [Kim95]. In this study 2 groups of users (MIS students) were trained in one of the model types and then asked to answer a set of questions addressed to syntactic and semantic aspects of a test model in 'their' type. A problem with this study is that the 2 data model types studied have similar graphic styles as defined below.

In the following, we present a study of the effect of graphic style on reader perception of data models. Three different graphic styles were represented in the experiment: *fully graphic*, *embedded structure*, and *list based.* The evaluation is based on the eye tracking and audio data taken while subjects interpreted a set of models expressed in the above graphic styles. The model types used were modified slightly to emphasize their basic graphic style. (See the Appendix for an illustration.)

1.1 Graphic representation of Data Model Concepts

The basic concepts of data models include *entities* (objects), inter-entity *relationships*, and *attributes,* (characteristics or roles of entities and relationships). *Methods* represent the behaviour of an entity. Graphic data models represent entity types as graph nodes and relationship types as connecting lines, possibly with a line-node symbol to which the name and characteristics of the relationship are associated. Several graphic data model types also include specification of attribute domains. Graphic data models can be classified according to their use of graphic symbols as: (see also Figure 1a)

1) *fully graphic*, where each model concept is represented by a specific graphic symbol, example: the NIAM model [Nijs89],

2) *embedded structure*, where entity and relationship concepts have specific graphic symbols within which attributes, with domain specification, and methods are listed, example: the object-oriented data model, OODM, [Catt91].

3) *list based*, where entity and relationship concepts have specific graphic symbols, to which lists of attributes, with their domain specification, are attached, example: the structural semantic model, SSM, [Nord93].

1.2 Constraint Representation

System constraints include *identification and link attributes (keys), relationship cardinalities, subclass specification, and domain value sets.* Representation of constraint specifications account for the greatest variation among graphic model styles. Constraints are commonly represented as annotations or special symbols placed in/on/or adjacent to the connection lines or the entity/relationship/attribute node symbols. For example, identification attributes can be underlined (SSM) or embedded within the entity node (NIAM), while restrictions on the domain value set are frequently an annotation to the domain specification. (See also appendix examples)

Cardinality constraints, giving required and maximal participation in a relationship, have the most varied representation. In the experiment model types, a mandatory *'dot*' notation is used in NIAM models, *arrow head* notation is used in NIAM and OODM models, while a *'(min:max)*' notation is used in both the NIAM and SSM models. Node notation is used for classification participation characteristics in the SSM, while the NIAM and OODM use a more traditional subclass notation. Figure 1b shows the graphic symbols used to represent a 1:10 cardinality for the relationship from E1 to E2 and total, disjoint participation specifications for the subentities SE1 and SE2[1].

New data model types, using variations of the above graphic styles and notations, are continuously being presented. Presumably authors of 'new' model styles assume that 'their' model style facilitates model legibility and user comprehension.

1) Note that the cardinality for the relationship from E2 to E1 is not defined in this figure.

2. GRAPHIC MODEL PERCEPTION - An Experiment

One method for gathering data to study the effect of graphic style on model perception is to track eye movements while subjects give an oral interpretation of models presented in various graphic styles. Eye movement registrations consist of fixations and saccades (movement between fixations), where a fixation is defined as the length of time the eye remains focused on one area. Processing information occurs only during fixations [Wolv83]. A minimum fixation duration of 100 to 125 milliseconds is needed to view an item and transmit that information to the brain [McCo83]. According to a reading model based on immediacy and eye-mind assumptions, text is interpreted immediately during an eye fixation, the fixated word is available for cognitive processing within a few tens of milliseconds and the eye stays fixated on a word as long as necessary to process it [Just80]. In graphic model perception, we were particularly interested in:

- Which graph components and symbols receive the most attention,
- Which graph components are not seen and which are seen but not reported.
- If graphic style effects model comprehension.

2.1 The Experiment

Data models of equivalent complexity were designed for six information systems chosen from the application areas: *education, project-management, library, concerts, sales,* and *purchasing*. Each of these systems was modelled using comparable versions of the NIAM, OODM, and SSM methodologies. (See the models for the *project* application given in the Appendix.) Each model contained 2 independent interrelated entities, a classification structure with 2 subentities, 24 attributes including atomic, composite, derived, and multivalued attributes, domain specifications with valid data value sets, primary key specifications, relationship cardinalities, and participation constraints. A fourth model type, IDEF1X [Loom86] was included in the experiment and used to determine the general data model interpretation ability of the subjects. Six experiment sequences were defined to compensate for placement of the model types.

Immediately proceeding the experiment, the subject was given an example data model and verbal explanations of the terminology used. The subject was informed that the experiment consisted of 8 models, 2 each of 4 different data model types, and instructed to give a complete, oral interpretation of each model and to indicate when each interpretation was finished. Maximum interpretation time allowed per model was 4 minutes. Average interpretation time per model was 2.25 minutes. The verbal protocol was recorded on tape.

Eye movement data was collected using an Applied Sciences Laboratory Eye Movement Monitor which was connected to a Macintosh II computer. An infrared beam projected at one of the subject's eyes was used to compute the location coordinates of the eye, 60 times a second as the subject viewed each graphic model on a large video monitor. A fixation was defined when at least 10 consecutive points were within a 10 by 18 pixel area, giving a fixation duration of 167 msecs. An area was defined for each component in the data models, such that all concepts and constraints had identifiable graph areas in each model.

2.2 The Subjects

System users, who are expected to read and confirm the correctness of data models, can be expected to have good knowledge of their application area (and thus the topic area for the data model), some knowledge of computer applications, and some familiarity with the types of models used with computer applications. However, it is not reasonable to expect them to have training in a particular data model type. Ideally, the subjects for an experiment to test data model perception should represent this user group.

The subjects in this experiment were 17 volunteer students from the Dept. of Information and Computer Science at the University of Hawaii at Manoa (ICS/UH): 12 senior-level undergraduates from a course in systems analysis and design, and 5 graduate students. Though they did have training in the development of system design models, and in particular in the use of the IDEF1X model which is taught and used at ICS/UH, none were familiar with the other models used in the experiment: NIAM, OODM, or SSM. It was assumed that the application areas used in the experiment were from a common domain of knowledge and thus familiar to the subjects. While students do not represent system users, it was assumed that their strategies for reading data models of familiar applications would be similar to those utilized by system users when reading a model of their application domain.

3 DATA MODEL COMPREHENSION

A *comprehension score* (Q) was calculated for each model interpretation as the percent of correctly identified model components. The Q scores for the 136 model interpretations varied from 25-95. Average scores, Q", were calculated from the Q-scores for each reader (variation from 34 to 77) and for each model-type (variations 52-71). In order to facilitate comparison with the model types which do not contain method assignment, the OODM Q scores were calculated without credit for method recognition.

3.1 Skilled vs Nonskilled Readers

The subjects were grouped into two skill levels; *skilled and non-skilled* graphic model interpreters, according to their average Q score for the IDEF1X models. Skilled IDEF1X model interpreters also demonstrated above average Q" scores for the NOS (NIAM, OODM, and SSM) model interpretations, while the nonskilled readers had below average NOS Q" scores. Table 1 shows the average comprehension scores for the interpretations of each model type for all subjects, and for the skilled and nonskilled groups.

Level	***IDEF1X***	***NIAM***	***OODM***	***SSM***
Total	71	52	59	58
Skilled	80	58	68	67
Nonskilled	61	46	49	48

Table 1: Average Comprehension scores for each Model Type for All Subjects and Skilled and Nonskilled Groups

As noted above, the readers were familiar with IDEF1X models. Of the unfamiliar NOS models, the NIAM models were least well understood. Skilled subjects had significantly[2] poorer scores for the NIAM model interpretations. One explanation could be that the subjects were unfamiliar with basic data modelling concepts. Another, that the subjects were familiar with the concepts but did not recognize their graphic representation.

3.2 Familiarity of Data Model Concepts

The experiment models contained the basic data modelling concepts: *entities (objects), attributes, relationships,* and *domain sets,* and the constraint types: *primary key, relationship cardinality, subclass participation,* and *data-value sets.* The OODM models also contained *method names*. Table 2a shows the percentage of readers who recognized each concept and constraint type in at least 1 model interpretation. Clearly, entities, attribute/ methods, and interentity structures were familiar concepts for this group of readers. The relatively low recognition of domain specifications may be attributed to a lack of expectancy since the familiar IDEF1X models do not include domain specifications. Table 2b shows the percentage of models of each type in which the concepts[3] and constraints were correctly identified. Note that recognition of the interentity relationships and their constraints is relatively poor in the OODM and NIAM models.

	Concept					***Constraint***			
	Ent	***Atr***	***Rel***	***Dom***	***Meth***	***Key***	***Car***	***Par***	***Val***
Total	100	100	100	76	88	76	94	100	41
Skilled	100	100	100	89	100	100	100	100	78
Nonskilled	100	100	100	63	75	50	88	100	0

Table 2a: Concept Recognition by Skill Level, percent

	Ent	*Atr*	*Rel*	*Dom*	*Meth*	*Key*	*Car*	*Par*	*Val*
IDEF1X	100	100	93	-	-	53	57	80	-
NIAM	100	100	70	60	-	13	17	13	17
OODM	100	97	63	23	77	3	3	3	20
SSM	100	100	90	40	-	10	30	67	30

Table 2b: Concept Recognition by DM Type, percent

It appears the choice of node symbol for entity representation and the placement of the attribute list does not effect entity and attribute recognition. The relationship symbols used in the NIAM and OODM models are not as distinguishable[4] as those of the IDEF1X and SSM models. In general, constraints were poorly identified indicating that graphic detail was not well recognized by these subjects. One explanation is that they were seen but not understood. A second explanation is that they literally were not seen.

2) p=0,054 on a paired two sample ANOVA test with significance set at 0,05.

3) The '-' indicates that the concept or constraint is not included in the model type.

4) The 'unrecognized' relationships were identified as entity types, leading to a certain confusion in model interpretation.

3.3 Interpretation Effort

Eye movement data, taken during scene interpretation, describes how a scene is read, which areas receive attention and which are not seen. Studies have shown that scene complexity influences the number of eye fixations used for interpretation. We assume that fixation count[5] is related to the effort expended in reading and interpreting the data models.

The average fixation counts for the model types was: IDEF1X: 173, SSM: 181, OODM: 205, and NIAM: 223. As anticipated, the familiar IDEF1X model was the 'easiest' to read. The SSM model required only %5 more reading effort, while the OODM and NIAM models required 18% and 29% more effort for interpretation, respectively. Table 3 shows the average percent of fixations used in each concept/constraint area in each model type. The method list in the OODM models is placed immediately following the attribute list. In the table, fixation counts for OODM methods (21%) have been added to the attribute count.

	Ent	*Atr+ Meth*	*Rel*	*Dom*	*Key*	*Car*	*Par*	*Val*
IDEF1X	8	43	11	-	24	5	9	-
NIAM	13	56	15	7	2	1	6	1
OODM	17	53	13	5	-	1	9	3
SSM	14	39	18	2	2	4	19	3

Table 3: Concept and Constraint Fixations per DM Type, percent

The basic entity-relationship structure was relatively quickly identified in all model types, requiring between 19-32% of the total fixation counts. The high percentage of fixation counts for attribute and method identification reflects the structure of the attribute areas where each contains a list of 3-6 attributes plus 0-3 methods for the OODM models. The SSM attribute list was most easily read, while more than 50% of model interpretation effort was spent identifying the NIAM and OODM attribute sets. Domain specifications were given little attention. This is curious considering that both the OODM and SSM models use a programming style which should be familiar to CS students.

Relatively little attention, 10-28% vs 38%, was given to identification of the constraints in the unfamiliar NIAM, OODM, and SSM models, compared to the familiar IDEF1X models. A notable exception is the SSM participation constraint, which alone received 19% of the interpretation effort. This was also 'rewarded' with a 67% correct interpretation rate (see tab.2b). It appears that there is a corelation between low focus and low interpretation which would support the explanation that the unarticulated concepts were simply 'not seen'.

5) The distribution of fixation duration within each graph area is equivalent to the fixation count distribution.

4. OBSERVATIONS

Graphic data models are intended to support user-designer communication during planning and specification of an information system. Usually, an IT trained system analyst develops a data model and asks the user to confirm that it is correct with respect to system requirements. A data model presents several aspects of information systems, including:

- the entity types about which information (data) is to be recorded by the system,
- the relationships, both associations and classification structures which record the activities and roles of the entities in the system, and
- the information system constraints which specify valid values and relationships.

We have tested the effect of 3 graphic styles, highly graphic (NIAM), embedded structure (OODM), and list structured (SSM), on data model comprehension.

4.1 The Effect of Graphic Style on User Comprehension

The subjects were generally familiar with the concepts and constraints represented in the experiment, with the skilled readers (as a consequence of definition) most familiar. While the interpretation of data models in unfamiliar graphic styles was generally not strong, skilled readers had significantly poorer interpretations of the highly graphic, and thus relatively dense, model types. Non-skilled readers identified correctly only the basic entity-attribute structures in the models, irrespective of graphic style.

4.2 The Effect of Graphic Style on Concept and Constraint Recognition

Both skilled and nonskilled users are able to identify the entity - attribute sets irrespective of the graphic symbol (oval vs rectangle) used for their representation. However, inter-entity relationships and classification structures, though known to the subjects, were more likely to be incorrectly interpreted when their representation symbol resembled that of the entity types.

Skilled readers saw more of the graphic detail, given as line and/or node annotation. Particularly, concepts which were placed in an 'open' graph area (SSM cardinalities) and/or had a separate graphic symbol (SSM participation specification) attracted attention and thereafter recognition.

The *Key Identification*[6] attributes in the experiment models were all named *Id*. While they were recognized in the familiar IDEF1X model by all of the skilled readers and half of the nonskilled readers, only 2 of the 9 skilled readers recognized key attributes in the other model types, and only one of these in all model types. Neither of the SSM underline nor the NIAM parenthesis' graphic strategies appears to support key recognition.

Relationship Cardinality Specifications, given as line-end *mandatory 'dot*' notation (NIAM) or *arrow head* notation (NIAM and OODM) were recognized in only 17% and 3% of these models, respectively, while the *(min:max)* notation, placed centrally on the relationship line, was recognized in 30% of the SSM models.

---

6) No graphic specification of key attributes is given in OODM models.

Participation Constraints, represented in separate structures (IDEF1X and SSM) were interpret correctly in 80% and 67% of these models respectively. However, the 'mathematical' set notation (NIAM and OODM) was recognized in only 13% of the models (NIAM).

Domain and Data Value Specifications are given in a programming style in both the OODM and SSM models:

OODM:	Attribute-name	: data-type (length | value list | value-range)
SSM:	Attribute-name	<data-type (length | value list | value-range) >

These were described by only 56% skilled readers. Two of the nonskilled readers (25%) recognized these specifications in the SSM model while none of this group articulated domains in the OODM models. This is interesting considering that all readers had some proficiency in programming and this form for data specification.

Domains are considered an object type in the NIAM and modelled with the same graphic symbol as entity types. This graphic strategy was successful in that 2/3 of the skilled readers and 50% of the nonskilled readers recognized these domain specifications.

4.3 The Effect of Graphic Style on Interpretation Effort

The list style model type (SSM) was most easily read requiring only 5% more effort than the familiar, embedded structure model type (IDEF1X). Interestingly, the embedded structure style model (OODM) required 18% more effort than the familiar model type which has the same basic graphic style. The highly graphic models (NIAM) required most effort, 29% more than the familiar, embedded structure model type.

Graphic detail, representing known system constraints, generally received little interpretation effort (10-28%). In particular: cardinality identification 'used' only 1% to 4% of the interpretation effort, participation identification 6% to19%, and domain specification 5% to 8%, as measured by percent of fixations.

Attention given to domain and data-value specifications was low, < 10%. It is possible that these readers, who were familiar with a data model type without domains, did not expect domain specifications in other data model types.

4.4 Summary and Further Studies

Our observations indicate that graphic style influences the level of detail seen in the model as well as the effort a user needs for interpretation, as measured by fixation counts. However, graphic style does not appear to effect identification of the general system structure, i.e. the entity types, attributes, and relationships which make up the system. While certain understanding can not be expected from a single experiment, we believe the observations summarized below are worth confirming.

- Highly graphic model types, exemplified here by the NIAM model type, appear to make a model more difficult to interpret than list based model types.
- Clearly separate symbols, in distinct graph areas, appear more legible than embedded symbols or line-end annotation.

It is disturbing that so many model interpretations omitted descriptive details, particularly the constraint specifications, which were given in the models and with which the interpreters were familiar. Why didn't they attract attention? Were they not understood? Were they considered unimportant? Supplementary experiments are planned to address these questions.

ACKNOWLEDGMENTS

A special thank you for help in preparation of the experiment, its execution, preparation of the data, and its analysis is extended to Professors W. Wesley Peterson and Svein Nordbotten. Also appreciated, is the interest and cooperation of Asst. Professor Donald DeRyke and the students of his systems analysis class.

REFERENCES

[Broo80] Brooke, J. and Duncan, K. An experimental study of flowcharts as an aid to the identification of procedural faults. Ergonomics, 23, 1980, 387-399.

[Catt91] Cattel,R.G.G. Object Data Management Object-oriented and Extended RDBS. Addison Wesley, 1991.

[Just80] Just, M., Carpenter, P., A theory of reading: From eye fixation to comprehension. Psychological Review, 87, 1980, pp 329 -354.

[Kim95] Kim,Y-G and S.T.March. Comparing Data Modeling Formalisms. CACM Vol.38:6, pp 103-115, 1995.

[Loom86] Loomis,M. Data Modeling - the IDEF1X Technique. Proc. IEEE Conf. on Computers and Communications. pp 146-151. March 1986.

[McCo83] McConkie, G. Eye movements and perception during reading. In K. Rayner (Ed.) Eye Movements in Reading: In Eye Movements in Reading: Perceptual and Language Processes, Academic Press, New York NY. 1983.

[Nijs89] Nijssen, G.M. and T.A. Halpin. Conceptual Schema And Relational Database Design - A fact oriented approach. Prentice Hall 1989.

[Nord93] Nordbotten, J.C. Modelling Relationships And Constraints In SSM - A Structural Semantic Data Model. Report no.14, ISSN 0803-6489. Information Science, Univ.of Bergen. 1993.

[Nord95] Nordbotten,J.C. and M.E.Crosby. Recognizing Graphic Detail - An Experiment in User Interpretation of Data Models. in Proc. BNCOD-13, Advances in Databases. Springer 1995.

[Petr95] Petre,M. Why Looking Isn't Always Seeing: Readership Skills and Graphical Programming. CACM Vol.38:6, pp 33-44, 1995.

[Shep82] Sheppard, S.B., Kruisi, E., Bailey, J. W. An empirical evaluation of software documentation formats. Proceedings Human Factors in Computer Systems, 1982, 121-124.

[Tore86] Teorey,T, D.Yang, and J.Fry. A logical Design methodology for Relational Databases using the Extended Entity-Relationship model. ACM Comp. Surveys, Vol.8:2, pp197-222, 1986.

[Wolv83] Wolverton, G.& Zola, D. The temporal characteristics of visual information extraction during reading. In K. Rayner (Ed.) Eye Movements in Reading: In Eye Movements in Reading: Perceptual and Language Processes, 1983, New York NY: Academic Press.

[Wrig73] Wright, P. & Reid, F. Written Information: Some Alternatives to Prose for Expressing the Outcome of Complex Contingencies. Journal of Applied Psychology 57, 1973, pp. 160-166.

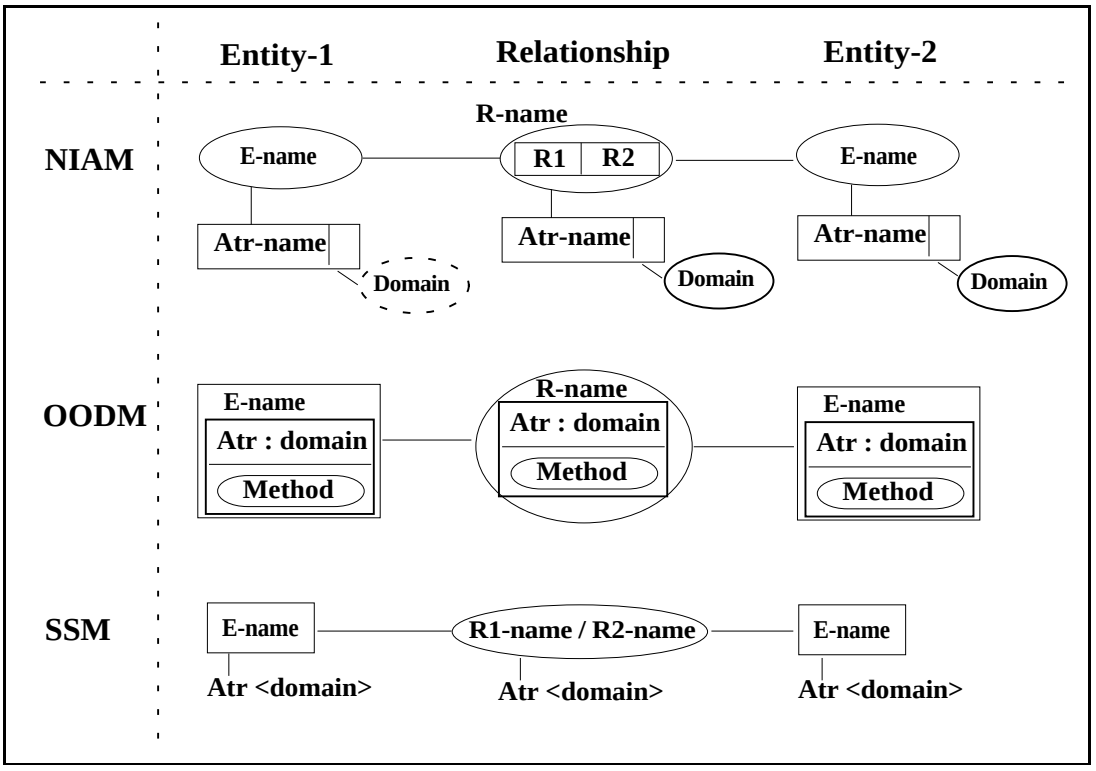


Figure 1.a: Graphic Representation of Basic Data Modelling Concepts

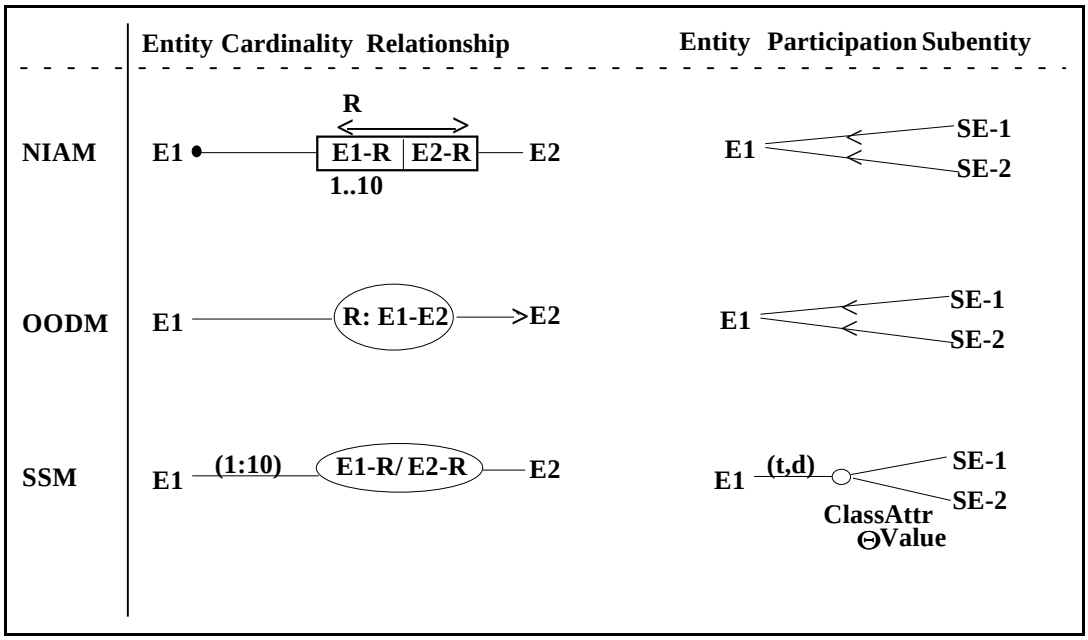
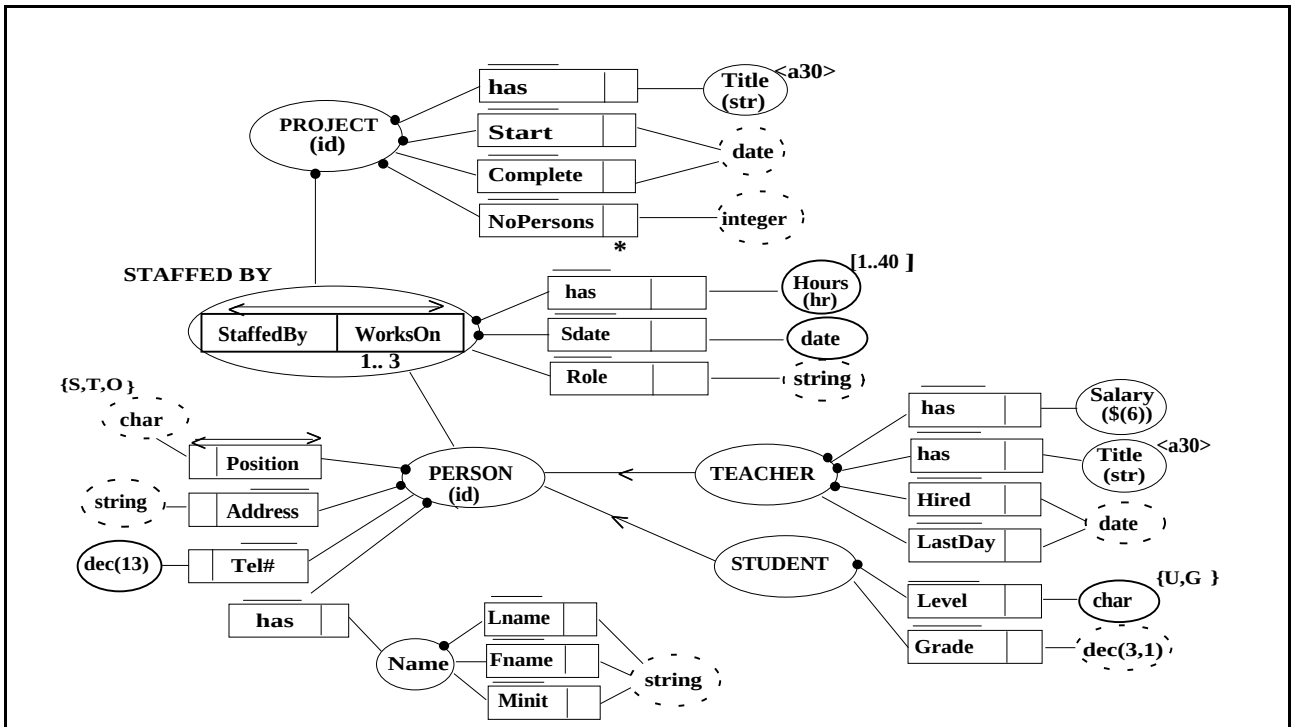


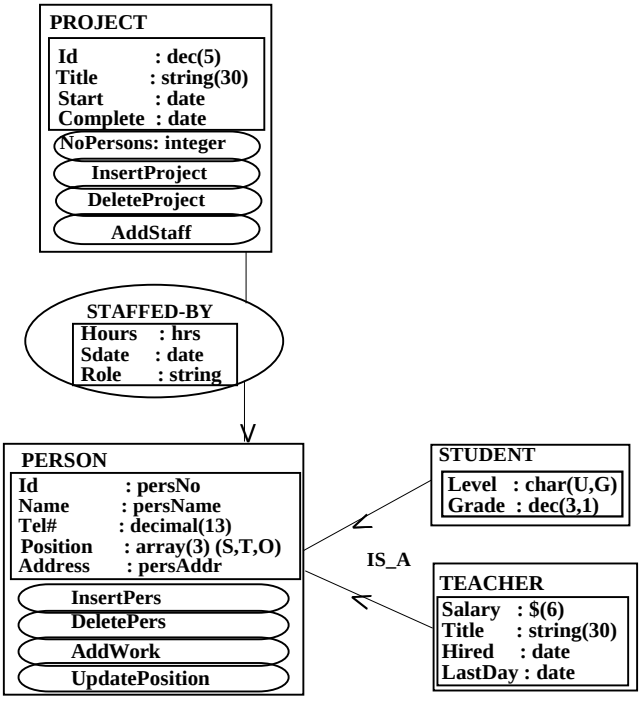
Figure 1.b: Graphic Representation of Relationship Constraints

APPENDIX: NIAM, OODM, and SSM Data Models for Project Management Application



NIAM

OODM



SSM