CICERO COLORING SCHEMA FOR DATA MODELING

Vladan Jovanovic Georgia Southern University, vladan@georgiasouthern.edu
Lily Cupic Georgia Southern University, lilycj48 @yahoo.com

ABSTRACT

Developing understandable data models as blueprints i.e. diagrammatic representations of large Databases and Data Warehouses is challenging, given the inherent complexity of large systems that DB/DW support. In this paper, we present a novel color-coding approach aimed at making large data models more understandable. The coloring schema (CICERO) is presented within an Ontology based approach to Data Modeling intended for a domain of inquiry comprised of system and problem situations primarily occurring in database and data warehouse design and integration. Paper also provides significant baseline data regarding defects in data modeling prior to using model coloring.

Keywords: Generic Models, Universal Models, Ontology-based Data Modeling

INTRODUCTION

The focus of our research is primarily on large data models and their use in teaching, irrespective of diagramming notations (UML, etc.). For this paper, we selected the IDEF1X standard notation [10] for illustration of our coloring approach to data model visualization. Furthermore, we define a large data model as a model with between 1,000 and 10,000 key terms to be defined – corresponding to individual concepts that is entities, attributes, and relationships. Subsequently we use orders of magnitude for classes of model size so that a small model applies if less than a 100 concepts is to be defined, a medium if less than one 1000. Similarly, very large models cover over 10,000 terms and extremely large models involve over 100,000 terms. For practical purposes of inquiry we do not expect coloring of a model to be effective beyond large size models. It is worth noting that, given above definitions of size, the great preponderance of models found in literature are small to medium.

A systematic use of color in professional object modeling presented in [1] significantly influenced our views. Nevertheless, it is our contention that a color coding of a data model, with an explicit systematic categorical foundation (we may say an ontology based approach underneath), is not only useful for modeling large systems but can be fairly effective in teaching modeling to novices and/or students. Study of complex industrial strength data modeling artifacts and knowledge transfer of lessons learned by experienced data modelers can both be helped with effective use of color. Prior art relevant for this work, in terms of modeling generalizations, come out of the pattern community, specifically works [2, 3, 4], works on Universal [5, 6, 38] i.e. standard data models, Generic data models [7, 8, 9,11], and widely respected data modeling, for example [13,14,15,16,17]; representative bibliographies can be found in [18], at DAMA website and from the authors.

The paper presents a coloring approach, with preliminary results of its use, and is organized as follows: a) a section on coloring schema, b) discussion of the experimental classroom use of the approach for comprehension/tailoring of large models such as [5,7, 9], and c) a short conclusion outlining expected directions of further study.

CICERO COLORING SCHEMA

A scheme (Figures 1-3) features a variation of the six base colors for stereotyped categories, and two light pastel colors for supplementary archetypes (meta categories). The CICERO schema is completed with white background intended only for domain specific events (processes and facts). To demonstrate the intuitiveness of our approach when dealing with large data models, we analyzed examples from [5, 7, 8, 11, 19-22, 27, 32]. The CICERO schema was evolved during experiments with a generic, that is, domain-independent approach to data modeling. Nowadays such work is openly considered an ontology based framework [38], but that was not clear when we started with it in 1982. During our work on large models it become clear that color-coding enhances the understandability of the meta-structural properties of models. In this regard, our approach has been particularly beneficial for students when dealing with larger models. Ideas for CICERO color selection are generally derived from [12] as well as from experiences with and limitations of ERwin.

The categories for which different colors are proposed are identified by numbers here only in order to help recognize diagram intent in lack and white printouts, as we do not recommend this usage. Color is an additional element of information guiding

reading (and understanding) of diagrams and numbers or color names will only interfere. Any further discussion of color palette, hue etc. in black and white will be quite inconvenient. The categories are numbered with general indication of the color used as follows:

1. Time (blue green)
2. Space (yellow)
3. Subject (light green)
4. Organization (darker green)
5. Object (red)
6. Convention (pink)
7. Role (gray)
8. Description (blue)
9 Event. (white).

<insert Figures 1-5 here>

The most versatile modeling mechanism is that of a ROLE covering multitude of functional specializations (for example: employer, customer, vendor, insurer, member, auditor, subcontractor, lender etc.), and effectively bridging and/or generalizing basic active categories (PERSON, SUBJECT, and ORGANIZATION). Similarly DESCRIPTION can be attached to pretty much everything to provide metadata and/or details. Model expansion (building) mechanisms are composition (relating types) and decomposition (expanding types). Methodological observation: understanding of a domain typically starts with enumeration of types where structural static properties are conveyed by specialization and by hierarchical decomposition represented by identifying relationships extending types within the same category. Furthermore 'laws of the domain' are expressed in direct links to applicable Conventions, Space and Time. Capture of the systems dynamics is later on accomplished piece-mail by relating Events with relevant categories, the most obvious meta-example is data warehouse star schema with a fact table connected to relevant dimensions. The generic framework from static analysis is to be completed with dynamics, anticipating desirable changes and that is where the skill of abstraction gets really tested and modelers earn their reputation. Individual experience and persuasiveness are challenged in the global environment where a shared meaning for a wide community (i.e. an ontology) is a precondition for successful systems development and integration.

CONCLUSIONS

Our expected directions of future study can be summarized as follows: detail analysis of large data models, such as those mentioned in [20-32], and a further formulation and refinement of a viable method to model and design databases and data warehouses within the context of generic categories aka dimensions in data warehousing, (colored using CICERO schema) per each respective application

BASELINE CHARACTERIZATION

We can now state a not so humble requirement to simplify presentation of larger examples and provide visual consistency enabling recognition and use of patterns in teaching and learning of data modeling, but in reality the CICDERO coloring schema appeared at first almost as an afterthought aimed at helping transferring lessons learned during decades of large scale data modeling, to students. Our students of data modeling were having hard time orienting in anything but trivial size system models. It is the primary claim of our approach that coloring helped students in model comprehension. We can offer only one simple numerical comparison here as a 'proof' and revisit that when more data points (only 9 cases involved some use of coloring) are obtained. Statistical data regarding normalized defect density, forming our baseline, without use colors, are obtained over a period of 25 years on over 750 cases involving close to four thousand individuals (good number of them professional information system developers), with a breakdown shown in Table1.

<insert Tables 1-3 here>

While we accumulated quite a rich data sets for the baseline (without using color) we do not have at this point statistics regarding cases using the color (except preliminary examples which are experimental in nature and may best be excluded in any future analysis). The only quantitative comparison offered in this paper is really indirect and imperfectly measured, it is related to the time to operationalize a data model, meaning to introduce domain, conceptualize, implement, test (in most of the cases with Oracle 7 and above), and answer some questions on a brief project review. Anecdotal evidence indicates that meaningful customization of large models was possible under comparable time limits (to the times estimated per baseline data sets, Table-1, dealing with small to medium size cases) with large size models explicitly based on mature i.e. standardized baseline models [5, 19- 32]; and it was a shared opinion among students and faculty that without the help of categorization and the CICERO coloring schema the tasks would have been untenable.

domain. Related to that is our desire to perform substantial statistical analysis of the effectiveness of the use of color in data modeling, by various audiences and under varying conditions (notations, tools, previous experiences etc. etc.). The timeframe for data collection, unless other researchers join will be about 5-10 years. We had not addressed ontological choices made regarding generic categories underpinning CICERO scheme, references [33-36] are of interest but the topic requires a separate treatment and much further study.

REFERENCES

1. P.Coad, E. Lefebvre, J. DeLuca "Java Modeling in Color with UML" Prentice Hall 1999,
2. M Jones, I. Song, *Dimensional modeling: identifying, classifying & applying patterns,* DOLAP'05, Bremen 29-38, 2005
3. D. Hay, "Data Model Patterns- Conventions of Thought", Dorset House1996
4. M. Fowler "Analysis Patterns- Reusable Object Models" Addison Wesley, 1997
5. L. Silverston "The Data Modeling Resource Book" Vol I and II , 2nd edition, J. Wiley, 2001,
6. D. Marco, M. Jennings "Universal Meta Data Models" J. Wiley 2004,
7. W. Scheer "Enterprise-Wide Data Modeling", Springer-Verlag, 1989,
8. W. Bumpus, et all "Common Information Model" J. Wiley 2001,
9. L. Sanders "Data Modeling" Boyd & Fraser 1995,
10. IEEE Std. 1320.2-1998. IEEE Standard for Conceptual Modeling Language Syntax and Semantics for IDEF1X. New York: IEEE, 1998,
11. J. Gessford "How to build Business-Wide Databases" J. Wiley 1991,
12. M. Stone, A Field Guide to Digital Color, AK Peters 2003,
13. T. Bruce, Designing Quality Databases with IDEF1X Information Model, Dorset House 1992,
14. G. Simsion, G. Witt, Data Modeling Essentials, 3ed. MK 2005,
15. S. Hoberman "Data Modeler's Workbench" J. Wiley 2002,
16. R. Baker "CASE*Method Entity Relationship Modeling" Addison Wesley 1989,
17. T. Teorey "Database Modeling and Design" Morgan Kaufmann 3ed. 1999,
18. G. Simsion "Data Modeling Theory and Practice" Technics Publications 2007,
19. ISO 15962-2, Integration of lifecycle data for process plant including oil and gas production facilities, Part-2 Data Model, 2003,
20. www.dtmf.org CIM V2.14,
21. SDI model, preloaded in Data Extend Semantic Integrator, Progress 2008,
22. SMEF DM 1_10, BBC Standard Media Exchange Framework, BBC 09.20.2004
23. Unified POS, Retail Standard Data Model, 2005 (for members only).
24. AFCEE 1997, *Technical report for the ERPIMS 2.0 database model of the Environmental Resources Program Information Management System (ERPIMS),* US Air Force Center for Environmental Excellence, http://www.afcee.brooks.af.mil/ms/irpread.htm
25. ANZMETA metadata standard http://www.anzlic.org.au/metaelem.htm
26. Data Model for Geology http://www.ned.dem.csiro.au/research/visualisation/DMGE/
27. Data Model for Environment, BC http://ilmbwww.gov.bc.ca/risc/pubs/other/corporatestandards/assets
28. Digital Geologic Map Data Model v 4.3, September 1999, http://geology.usgs.gov/dm/)
29. Federal Enterprise Architecture Program: The Data Reference Model Version 2.0 November 2005,
30. Huntley, R. Curtis, T. A Standard PPDM Data Model for the Energy Industry, CSEQ Recorder November 2001 (for members only)
31. Miller, D.R., Hume, R.G. & Parker, A.J. (1997), The Geoscience Data Model P431. AMIRA Interim Final Report, Australian Mineral Industries Research Association
32. Public Petroleum Data Model (PPDM), http://www.ppdm.org .
33. T. Gruber A Translation Approach to Portable Ontologies, Knowledge Acquisition V5, pp199-220, 1993,
34. R. Hirscheim, H.Klien, A. Lyytinen "Information Systems Development and Data Modeling: Conceptual and Philosophical Foundations" Cambridge University Press 1995,
35. M. Heller The Ontology of Physical Objects, Cambridge University Press 1990,
36. T. Sider "Four-Dimensionalism: An Ontology of Persistence and Time", Oxford University Press 2001,
37. R. Meersman, P. Spyns, M. Jarrar "Data modeling versus Ontology engineering' ACM SIGMOD Record V31, N4, 2002,
38. M.West , C.Partridge, M. Lycett, Enterprise Data Modeling: Developing and Ontology-based Framework for the Shell Downstream Business" recently retrieved from www.

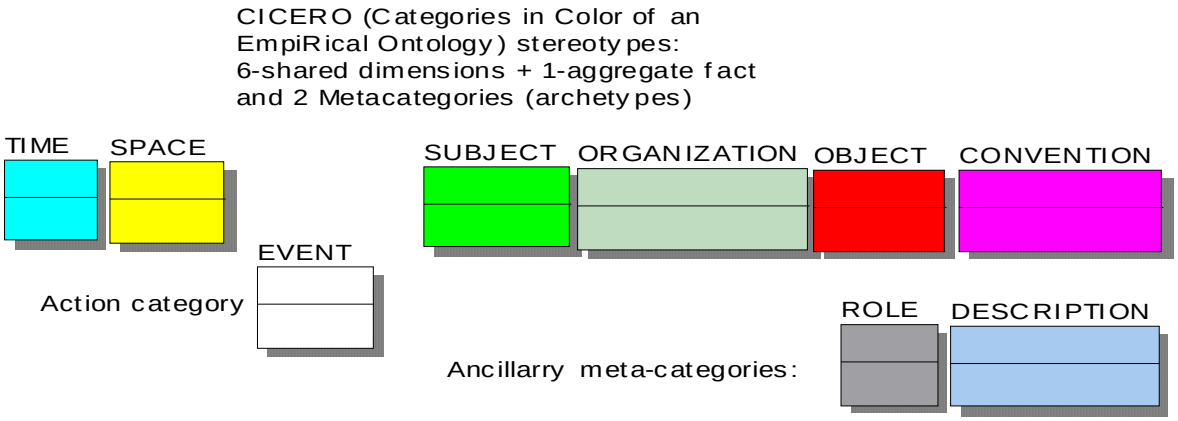

Figure 1. Categories in CICERO1

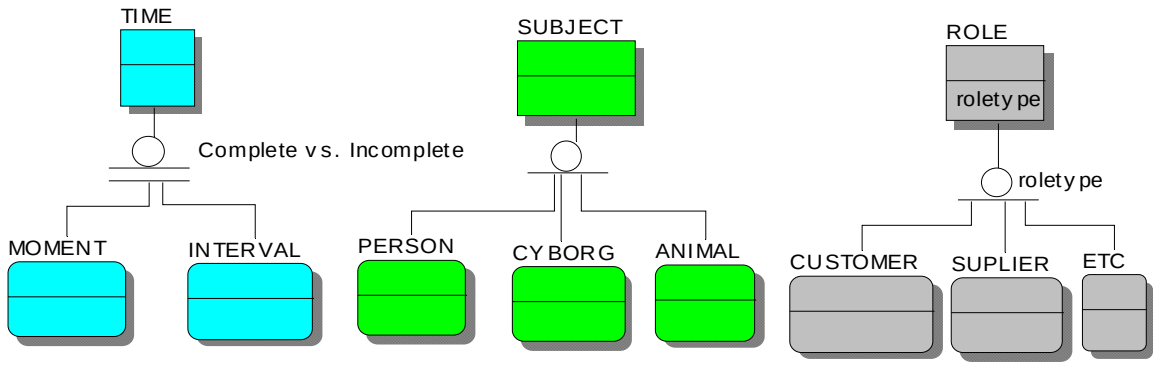

Figure 2. Example Specializations

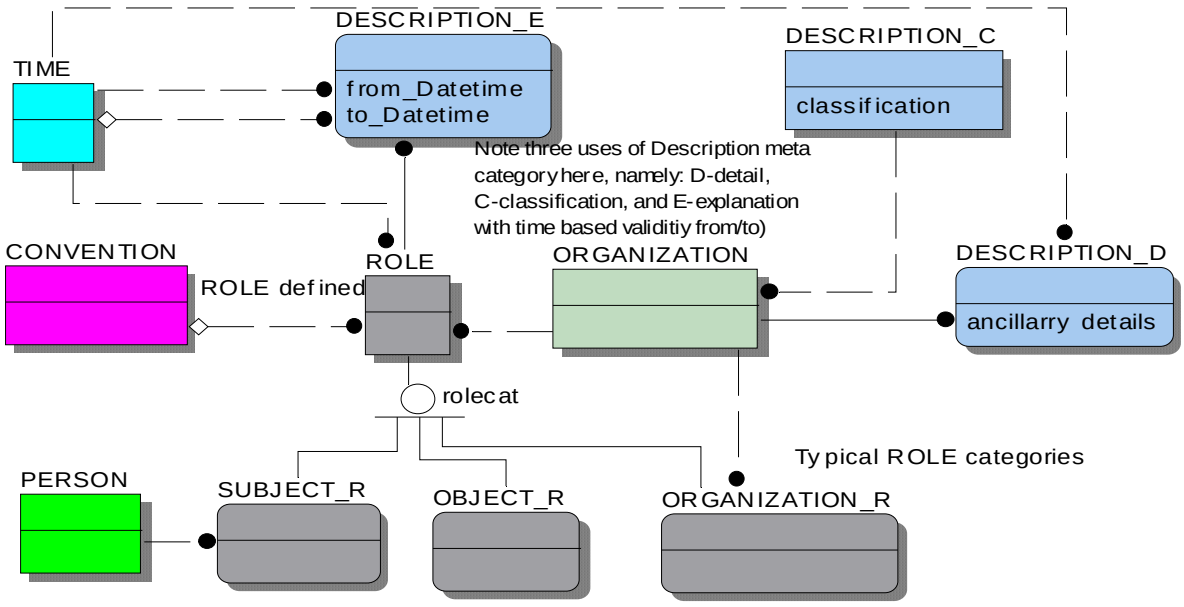

Figure 3. Description and Role

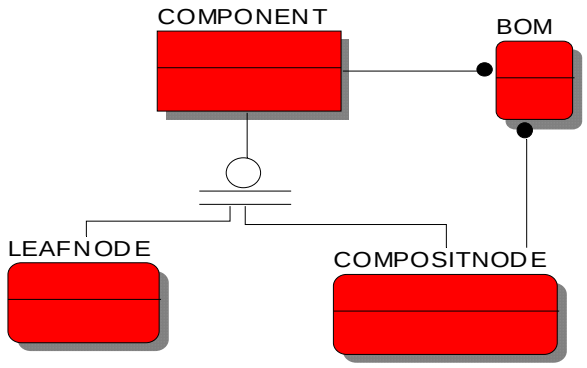

Figure 4. General Structure: Composite

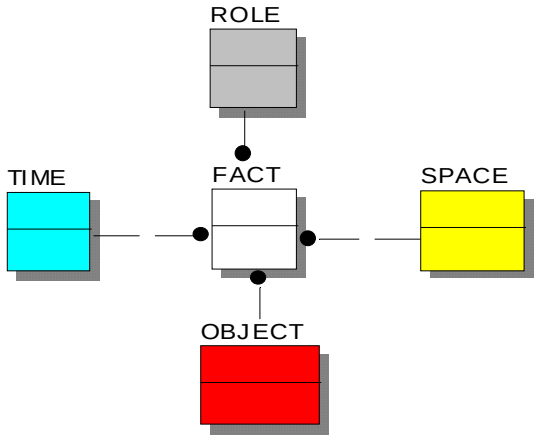

Figure 5. Process Fact in focus

Table 1. Field data sets from our practice, average time estimate for >90% cases is 3-5 weeks

Industry Workshops	Industry Design Reviews	Undergraduate Student projects	Graduate Student projects	Summary per Notation Used
Number of cases= 41 Avg. model size= 106 Defect Density= 26%	Number of cases=26 Avg. model size=288 D. Density=21%	Number of cases=64 Avg. model size=45 Defect Density=37%	Number of cases=34 Avg. model size=61 Defect Density=34%	Chen ERA Cases=165 Size [45-288] Defects [21-37%]
Number of cases=54 Avg. model size=78 Defect Density=21	Number of cases=16 Avg. model size=128 D. Density=19	Number of cases=208 Avg. model size=69 Defect Density=28	Number of cases=189 Avg. model size=76 Defect Density=23	Idef1X Cases=467 Size [59-128] Defects [19-28%]
Number of cases= 9 Avg. model size=77 Defect Density=27	NA	Number of cases=21 Avg. model size=44 Defect Density=39	Number of cases=1 model size=67 Defect Density=31	Oracle's CASE Cases=41 Size [44-77] Defects [27-39%]
NA	NA	Number of cases=25 Avg. model size=56 Defect Density=41	Number of cases=69 Avg. model size=102 Defect Density=35	UML Cases=94 Size [56-102] Defects [33-41%]
Cases= 104 Avg. size= [77-106] Defects [21-27%]	Cases=42 Avg. size= [128-288] Defects[19-21%]	Cases=318 Avg. size=[44-69] Defects=[28-41%]	Cases=293 Avg. size=61-102] Defects=[31-35%]	Total: Cases=757 Size= [44-288] Defects=[19-41%]

Table 2 Exploratory data using CICERO coding (defects data incomplete)

Industry Model Review	Undergraduate Students	Graduate Students	Notation Used
cases=1; size > 2000 Time~ 4 Weeks	cases=8; size>1000 Time~ 6 Weeks	cases=1; size >1000 Time ~5Weeks	Idef1X
cases=1; l size > 1000 Time~ 3 Weeks	cases=4; size >1000 time ~ 6 Weeks	NA	Oracle's CASE
cases=1 ; size > 2500 Time~ 5 Weeks	NA	NA	UML

Table 3. Types of Defects for Different Notations (stabilized version since 2005)

UML Class Diagram	**IIDEF1X ERA (with CICERO)**	**Chen's ERA Variations/ORACLE**
Missing a class	Missing en entity	Missing en entity
Missing an attribute	Missing an attribute	Missing an attribute
Missing an association	Missing a relationship	Missing a relationship
Needless class	Needless entity	Needless entity
Needless attribute	Needless attribute	Needless attribute
Incorrect multiplicities	Incorrect multiplicity	Incorrect multiplicity
Wrong class name	Wrong entity name	Wrong entity name
Wrong association name	Wrong relationship/role name	Wrong relationship/role name
Needless associations	Needless relationship	Needless relationship
aggregation/composition	NA	NA
Incorrect cardinality limits	Incorrect optional/mandatory relationship	*Incorrect rel exclusivity or optionally
	Wrong relationship type	Wrong relationship type
Incorrect inheritance	Incorrect supertype/subtype	Incorrect supertype/subtype
	Missing subtype/supertype	Missing subtype/supertype
Wrong attribute name	Wrong attribute name	Wrong attribute name
	Wrong std domain/attribute used	Wrong std domain/attribute used
	Missing standardized domain/attribute	Missing standardized domain/attribute
Wrong type	Wrong definition of std att-domain	Wrong definition of std att- domain
Missing 'Domain' Area	Missing Subject Area	Missing Subject Area
Incorrect Domain/package	Incorrect Subject Area	Incorrect Subject Area (submodel)
Needless subject Area	Needless subject Area	Needless subject Area
Confusing Diagram Layout	Confusing Diagram Layout	Confusing Diagram Layout
	Wrong categorization/color assigned	
Wrong text note	Wrong text note	Wrong text note
Diagram not attributed	Diagram not attributed	Diagram not attributed