The Data Organization

1251 Yosemite Way
Hayward, CA 94545
(510) 303-8868
info@thedataorg.com

Entity Analysis
And
Entity Patterns

By Rainer Schoenrank

Data Warehouse Consultant

Jul 2018

Biography

Rainer Schoenrank is the senior data warehouse consultant for The Data Organization. He has degrees in physics and computer science from the University of Victoria in British Columbia, and computer science California State University East Bay in Hayward, California. He has built data warehouses for clients such as Pacific Bell, Genentech, GE Leasing, SGI, PPFA, Brobeck, Bank of America, Clorox, Leapfrog and Intuitive Surgical. He can be reached at rschoenrank@computer.org.

Table of Contents

1. INTRODUCTION

During the database development process, as we move from analyzing the data and the conceptual data model to specifying the logical data model, we have as our main input the conceptual data model document.

The document contains:

- Entity List – names of the entities identified during the analysis process
- Entity Description – description of the data concept that the entity represents
- Conceptual Data Model Diagram – that looks a lot like Kimball's dimensional model with all the entities and the relationships between the entities and the process measurements.

To develop the logical data model, we need to repeat the same analysis process for each entity in the conceptual data model. As we get the list of attributes for the entity in the conceptual data model, we notice that there are patterns in the structure of the entities that are common across all the entities. These patterns are the result of the answers that the business gives to the questions concerning the nature of the entity's attributes and how the business uses the data in those attributes. In this paper, we will formulate the attribute questionnaire and analyze the effects that the possible answers have on the structure of the entity.

2. ENTITY SPECIFICATION

In the conceptual data model document, there is the list of attributes. For the logical data model we need to complete the attribute definitions by defining the following attribute properties:

- **Entity name** – the name of the entity to which the attribute belongs. This is one of the entities listed in the conceptual data model document, e.g., Customer
- **Attribute name** – the name for the attribute in the metadata dictionary. This is a string that is a unique value in the dictionary, e.g., Customer Name
- **Attribute meaning** – the description of the data concept embodied in the attribute. This is a paragraph giving the meaning of the attribute. The meaning is not duplicated in the dictionary, e.g., the name used by the database owner to label a customer.
- **Logical data type** – the name of the logical data type used to represent the data. See the list of logical data types. In this example, Customer Name is of type Name.
- **Default value** – the value assigned to the attribute by the Database Management System (DBMS) when no value is given. The types of default values are DBMS generated, application generated and no default is allowed.

This information forms the basis for the metadata dictionary.

3. ATTRIBUTE ANALYSIS

For each attribute of each entity in the conceptual data model document, there is a sequence of questions that must be asked and answered about the entity's attributes. The answers will determine the logical data model pattern required for the entity.

3.1 First Question - Attribute Purpose

The first question asks: What is the purpose of the attribute?

An attribute should not have more than one purpose. For example, the attribute address can be either a place where mail is delivered or a physical geographic location. The address attribute should not attempt to do both.

The purpose of the attribute can be one of:

1. Does it describe the entity? i.e., is it the entity's name, color, address, etc.?
2. Does it organize the entity by classifying it into a taxonomy? An attribute named GL Account organizes this sales transaction occurrence into the GL reporting hierarchy. The description of the sales transaction does not depend on whether or not there is a GL reporting hierarchy.
3. Does it establish a relationship between
   i. this entity and itself (reflexive relationship). An attribute named parent or spouse attempts to create a relationship between two occurrences of person.
   ii. this entity and the process measurement entity or other master data entities in the conceptual data model?

Since there are many relationships between entity occurrences, many ways to organize entity occurrences and many attributes that describe an entity occurrence, the conceptual model of the entity pattern that is created by the answers to Question 1 is shown in Figure 1.

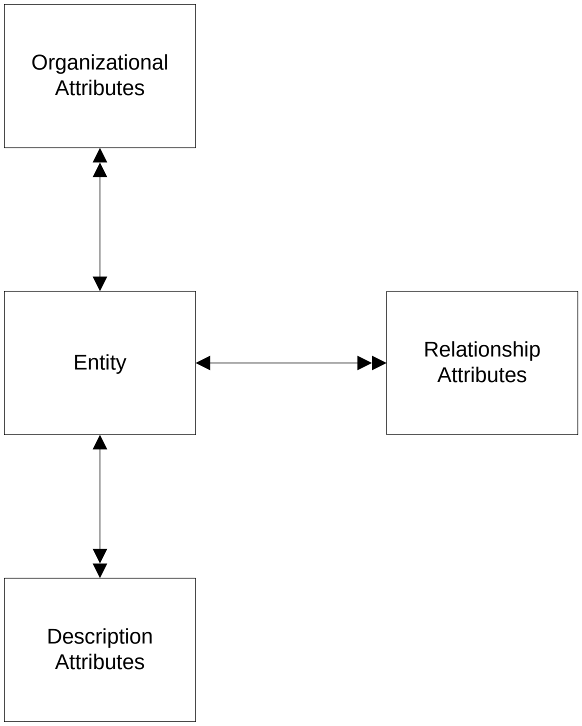


Figure 1. General Entity Pattern

The reason the organizational attributes are separate from the entity is because the organization of the entity does not impact the nature of the entity. The customer does not change because they belong to a particular sales team. The same is true of the relationship attributes. The relationships that a customer has with other customers does not alter the data required to become a customer. Whether the description attributes are placed into an attributive entity or are part of the base entity depends on the answers to the next two questions, but for master data management, all the description attributes have this structure.

3.2 Second Question

The second question that we ask of an attribute depends on the answer that we got from the first question. The three second questions are:

1. Is purpose of the attribute organizational? When the attribute is organizational then the next question is: Does the attribute place the entity into a taxonomy or does the attribute capture a decision of a group that the entity belongs to? The GL Account attribute or the Company Org Chart attribute place the entity into a taxonomy.

   The Sales Account attribute captures whether the Customer is a national, regional, or local account. This is not a description of the Customer rather it is the Sales Department's profile of the Customer. The Customer does not change because of what the Sales Department thinks.

2. Is purpose of the attribute a relationship? When the attribute is a relationship, then the next question is: Does the attribute create a relationship to another occurrence of the same entity or does it create a relationship to the sales process measurement entity. The relationship to the sales process measurement entity is part of the logical data model, but the reflexive relationship is like the Bill of Materials relationship for Product.

   You need to be very careful of reflexive relationships because of the amount of work involved to keep the data up to date. To relate a Customer to a parent Customer is very difficult in an active Merge and Acquisition business environment and it is not something required for the Customer to place an order.

3. Is purpose of the attribute descriptive? When the attribute is descriptive, then the next question is: Does the attribute describe the entity for all time or does the attribute describe the entity for a fixed period of time? This question is the basis of Kimball's dimensional types.

   The first type of attribute is the column like customer name. We expect the customer to have only one name and the value applies to the customer for all time. The second type of attribute is called a time dependent attribute (Kimball type II) that includes a begin date and an end date. The processing of this type of attribute is described as time interval processing.

3.3 Third Question - Attribute Occurrences

The third question to ask of an attribute is: Can the entity have more than one occurrence of the attribute?

For example, Customer identifier has a logical data type of identification. Does a customer have only one identification or does a customer have many identifications?

For an identification, the customer may offer a SSN, an EIN, a driver’s license, a passport, and so on. A customer could have many identifications, but our business may have a policy that chooses to use only one. This means our business will turn away customers that cannot be identified by the criteria we have chosen and this implies rework to the data model when our policy changes.

Another question about the attribute that should be asked and this applies to all attributes and does not change the entity structure solution is: does this attribute apply to all the subtypes of the entity?

3.4 Attribute Question Chart

The answers to the analysis questions are not inherent in the entity or its attributes, but are choices that the business makes about its data collection. The attribute question chart shows how the answers are generated and how each answer is related to a different entity pattern segment (bottom row of the chart). The white boxes in the chart show the generally accepted data modeling patterns while the blue boxes show the new patterns necessary to handle the answers to the analysis questions.

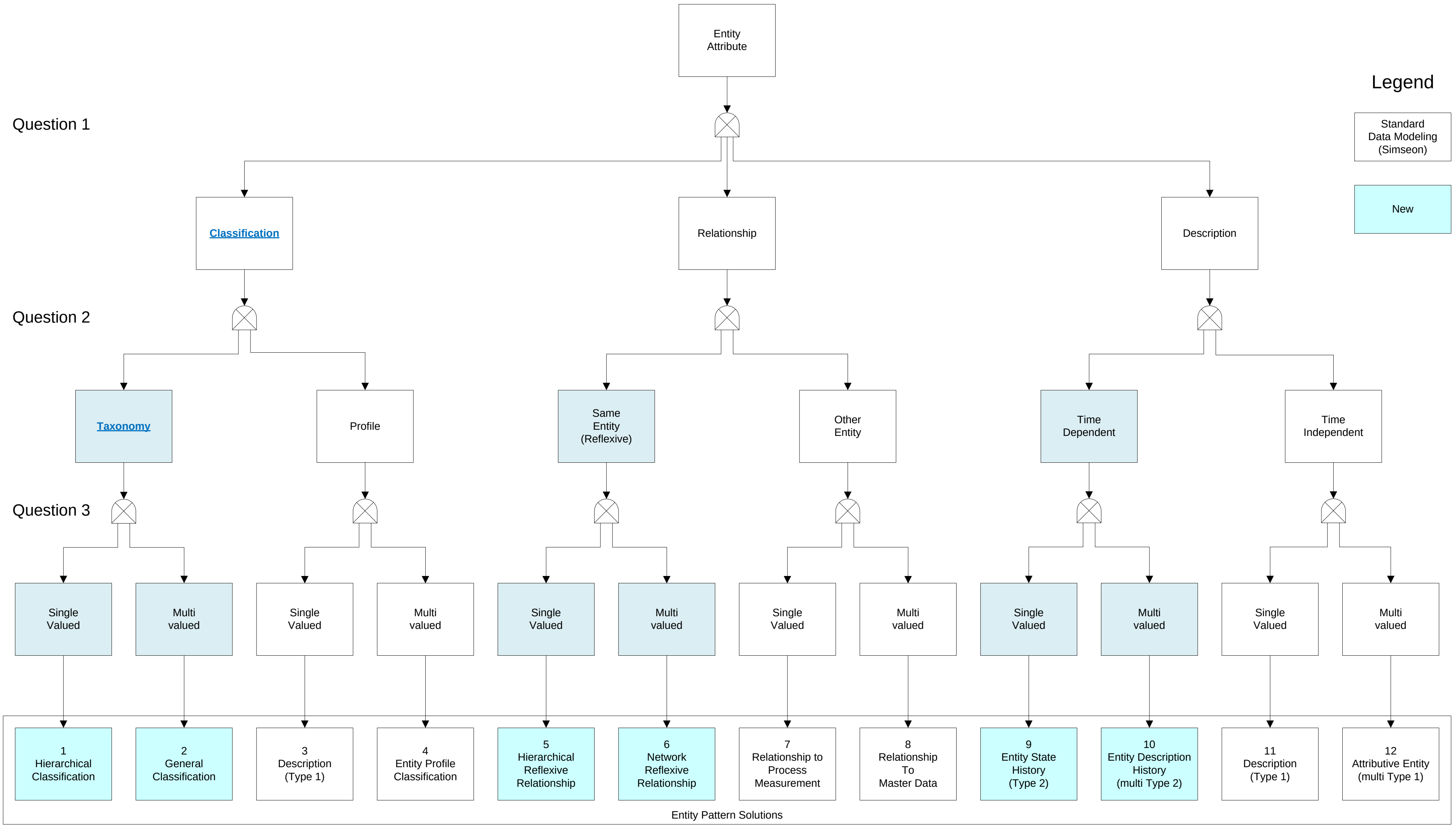


Figure 2. Attribute Question Chart

For an entity with 13 attributes, there are over 23 trillion different possible answers to the analysis questions and a similarly large number of detailed entity patterns. This means that the probability is negligible that two data models of Customer (i.e., two different application systems) will be using the same detailed entity pattern. They will have asked the analysis questions (or not) and come to completely different understandings of Customer. This is what makes the ETL portion of the data life cycle so difficult and tedious.

4. DETAILED ENTITY PATTERN

The detailed entity pattern diagram (Figure 3) shows the general case entity solution for the answers to the analysis questions. The entities from the conceptual data model are shown in blue and the answers to the analysis questions are shown in white with the arrows showing the relationships between the entity and the attributive entities.

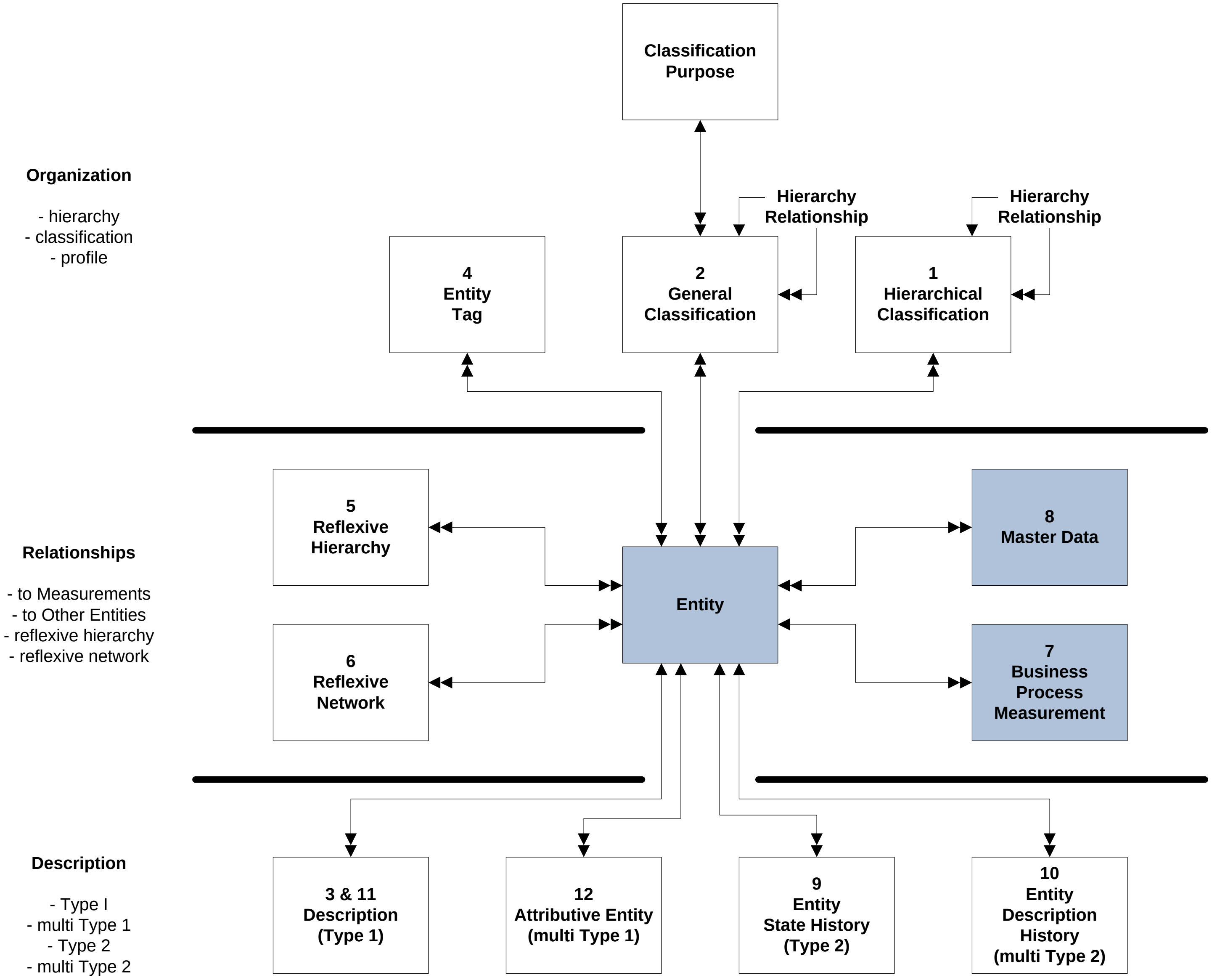


Figure 3. Detailed Entity Pattern

The description of the entity pattern solution is:

1. Hierarchical Classification – this attribute represents a special case of pattern 2. An example of a single organization of the entity is assuming that a business has only a single organization chart. The table labeled hierarchical classification will have the purpose of the classification in its name, e.g., Business Organization Chart Hierarchy. Having the purpose in the entity label violates the Information Principle (Codd) and is an error in the data modeling process. The entity attributes that are of this type should be generalized and included in the general classification type.
2. General Classification – the pattern for this attribute is described in the classification processing document
3. Description (Type 1) – implemented as a simple attribute for an entity
4. Entity Profile Classification – this attribute is a set of tags or categories for the entity and is implemented as an attributive table.
5. Hierarchical Reflexive Relationship – the pattern for this attribute is described in the reflexive relationships document
6. Network Reflexive Relationship – the pattern for this attribute is described in the reflexive relationships document
7. Relationship to Process Measurement – this relationship is shown in the conceptual data model and is implemented as foreign key column(s) in the measurement entity table.
8. Relationship to Master Data – this relationship occurs between two master data entities, e.g., the employee position history and is implemented as an associative table.
9. Entity State History (Type 2) – the pattern for this attribute is described in the time interval processing document
10. Entity Description History (Multi Type 2) – – the pattern for this attribute is described in the time interval processing document
11. Description (Type 1) – implemented as a simple attribute for an entity
12. Attributive Entity (Multi Type 1) – this attribute represents a descriptive attribute that has many values depending on the attribute type, for example, phone number. A customer may have an office phone number, a receiving phone number, a shipping phone number, etc.

5. USING THE ATTRIBUTES TO CREATE DATA MODELS

The entity analysis that documents the attribute assumptions is very powerful. There is one note of caution in this analysis. During the implementation and testing of the database, there will be suggestions to de-normalize the database. This request is usually based on the performance of the DBMS.

In terms of the entity pattern, this request means that the answers to the questions are not correct and the data model requires a redesign. Also, the de-normalization will create a tradeoff in the speed of database interface functions. The read functions will be faster because there are fewer database joins, but the update functions will be slower because there are multiple database rows that need to have their data modified.

In terms of the database implementation, the request means a reexamination of the database interface, the hardware tradeoffs that have been made and the primary purpose of the database (data capture or data reporting).

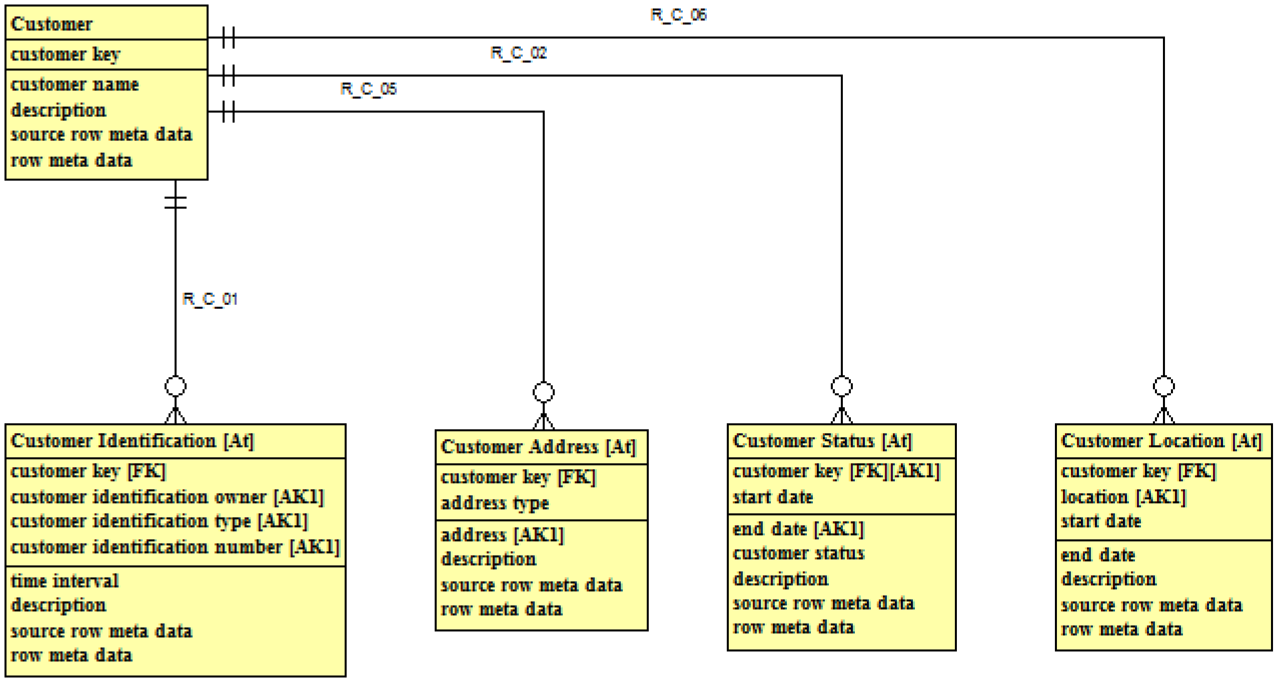
Using the Customer entity from the conceptual model document (on page 26), you need to prepare three worksheets that that identify the attributes, show the answers to the entity pattern questions and relate them to the entity pattern solutions. The examples are shown below.

5.1 Descriptive Attributes

The worksheet shows a sample list of the descriptive attributes. Other businesses may have more and different attributes and answers.

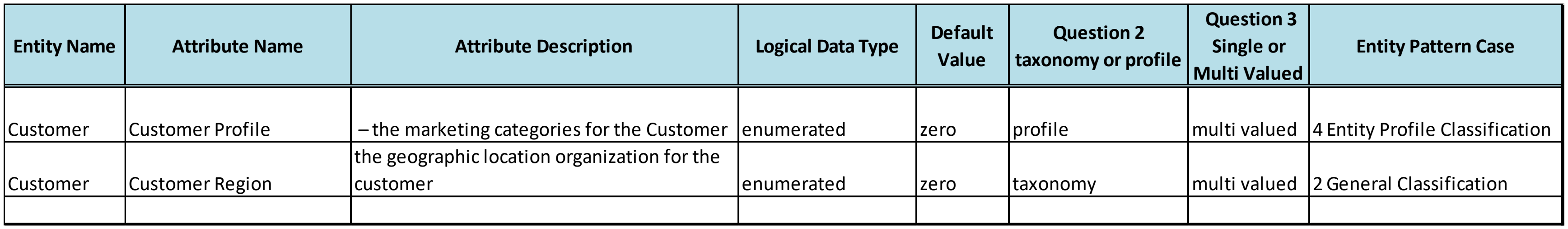
Entity Name	Attribute Name	Attribute Meaning	Logical Data Type	Default Value	Question 2 Time Dependency	Question 3 Single or Multi Valued	Entity Pattern Case
Customer	Customer Name	the labels that the customer is known by	name	null	constant in time	single valued	11 Description
Customer	Customer Identification	– the numbers that identify the Customer, i.e., IRS number, bank account number, credit card number, export permit number, etc.	identification	null	constant in time	multi valued	12 Attributive Entity
Customer	Customer Address	– the postal address for the Customer	address	null	constant in time	multi valued	12 Attributive Entity
Customer	Customer Contact	– the people who can be contacted about the state of the Customer	person	null	constant in time	multi valued	12 Attributive Entity
Customer	Customer Contact Email	– the Customer contact’s email addresses	structure	null	constant in time	multi valued	12 Attributive Entity
Customer	Customer Contact Phone	– the Customer contact’s phone numbers	phone	null	constant in time	multi valued	12 Attributive Entity
Customer	Customer Website	the names and locations of the Customer's websites	structure	null	constant in time	multi valued	12 Attributive Entity
Customer	Customer Authorization	– the data used to access the Customer	structure	null	constant in time	multi valued	12 Attributive Entity
Customer	Customer Balance	– the monthly balance of the customer's orders and payments	monetary amount	zero	changes in time	single valued	9 Entity State History
Customer	Customer Credit Line	the credit line available for an order	monetary amount	zero	changes in time	single valued	9 Entity State History
Customer	Customer Bill Info	– the billing cycle and payment method for the Customer	structure	null	changes in time	single valued	9 Entity State History
Customer	Customer Location	– the geographic location of the Customer	geographic location	null	changes in time	multi valued	10 Entity Description History
Customer	Customer Status	– the history of the Customer’s status in the sales process	enumerated	unknown	changes in time	single valued	9 Entity State History

The worksheet results would result in the partial customer descriptive data model (not all attributes are included) shown below.


Customer
customer key
customer name
description
source row meta data
row meta data
R_C_06
R_C_02
R_C_05
R_C_01
Customer Identification [At]
customer key [FK]
customer identification owner [AK1]
customer identification type [AK1]
customer identification number [AK1]
time interval
description
source row meta data
row meta data
Customer Address [At]
customer key [FK]
address type
address [AK1]
description
source row meta data
row meta data
Customer Status [At]
customer key [FK][AK1]
start date
end date [AK1]
customer status
description
source row meta data
row meta data
Customer Location [At]
customer key [FK]
location [AK1]
start date
end date
description
source row meta data
row meta data

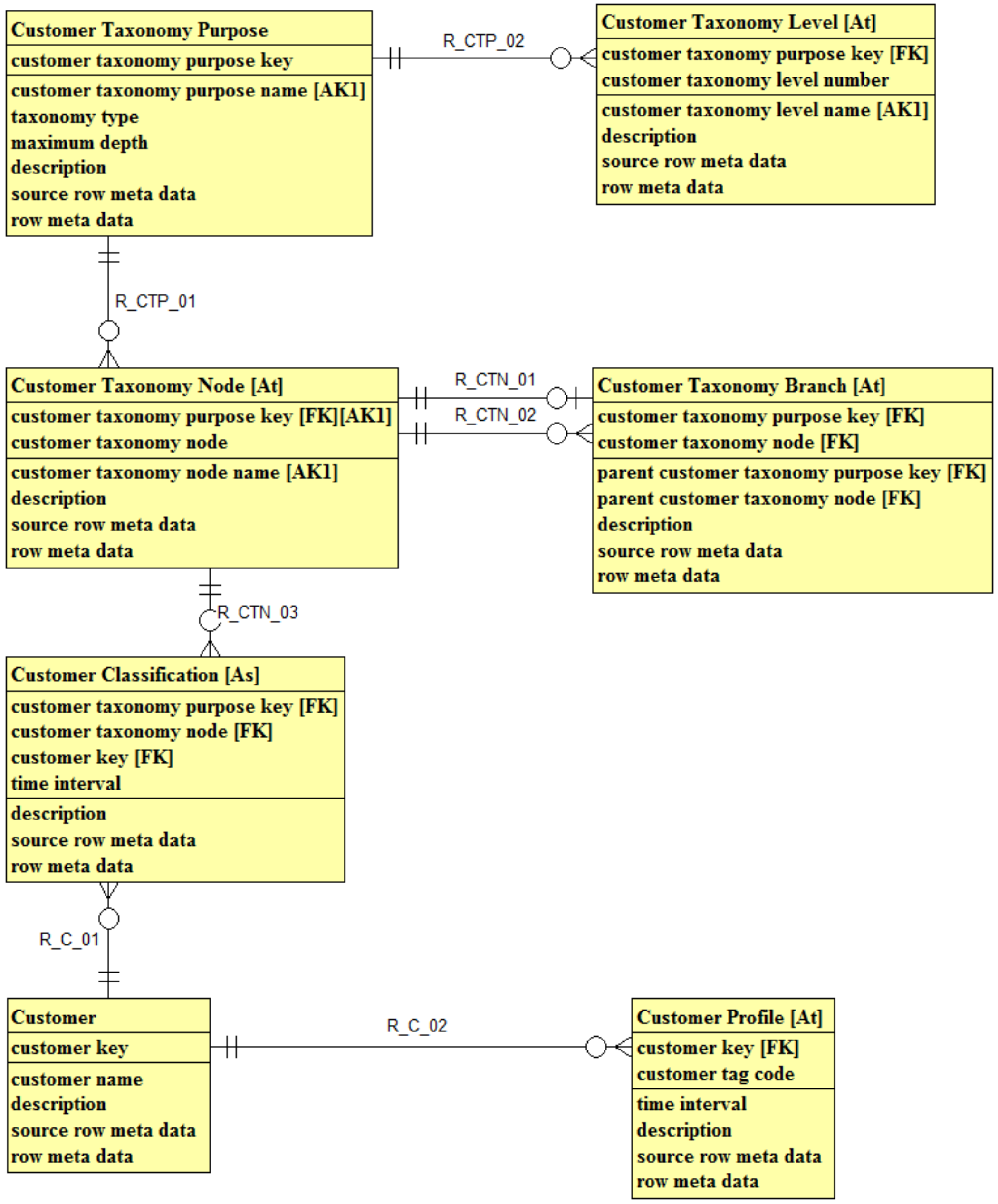
5.2 Organizational Attributes

The worksheet shows a sample list of the organizational attributes. Other businesses may have more and different attributes and answers.



Entity Name	Attribute Name	Attribute Description	Logical Data Type	Default Value	Question 2 taxonomy or profile	Question 3 Single or Multi Valued	Entity Pattern Case
Customer	Customer Profile	– the marketing categories for the Customer	enumerated	zero	profile	multi valued	4 Entity Profile Classification
Customer	Customer Region	the geographic location organization for the customer	enumerated	zero	taxonomy	multi valued	2 General Classification

The worksheet results would result in the customer organizational data model shown below.

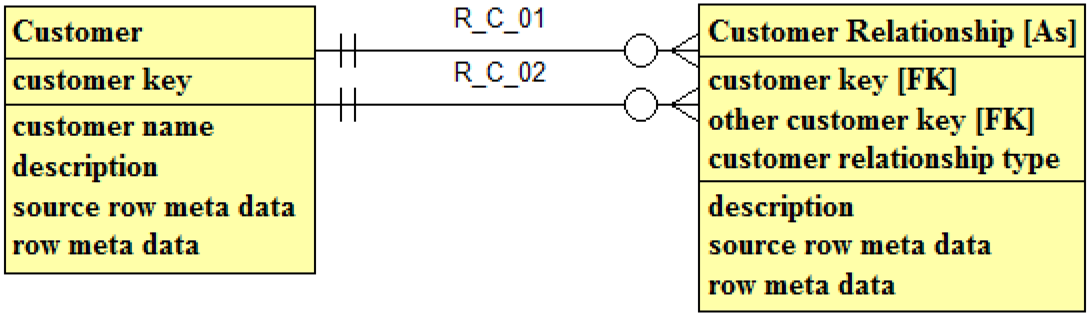

Customer Taxonomy Purpose
customer taxonomy purpose key
customer taxonomy purpose name [AK1]
taxonomy type
maximum depth
description
source row meta data
row meta data
R_CTP_02
Customer Taxonomy Level [At]
customer taxonomy purpose key [FK]
customer taxonomy level number
customer taxonomy level name [AK1]
description
source row meta data
row meta data
R_CTP_01
Customer Taxonomy Node [At]
customer taxonomy purpose key [FK][AK1]
customer taxonomy node
customer taxonomy node name [AK1]
description
source row meta data
row meta data
R_CTN_01
R_CTN_02
Customer Taxonomy Branch [At]
customer taxonomy purpose key [FK]
customer taxonomy node [FK]
parent customer taxonomy purpose key [FK]
parent customer taxonomy node [FK]
description
source row meta data
row meta data
R_CTN_03
Customer Classification [As]
customer taxonomy purpose key [FK]
customer taxonomy node [FK]
customer key [FK]
time interval
description
source row meta data
row meta data
R_C_01
Customer
customer key
customer name
description
source row meta data
row meta data
R_C_02
Customer Profile [At]
customer key [FK]
customer tag code
time interval
description
source row meta data
row meta data

5.3 Relationship Attributes

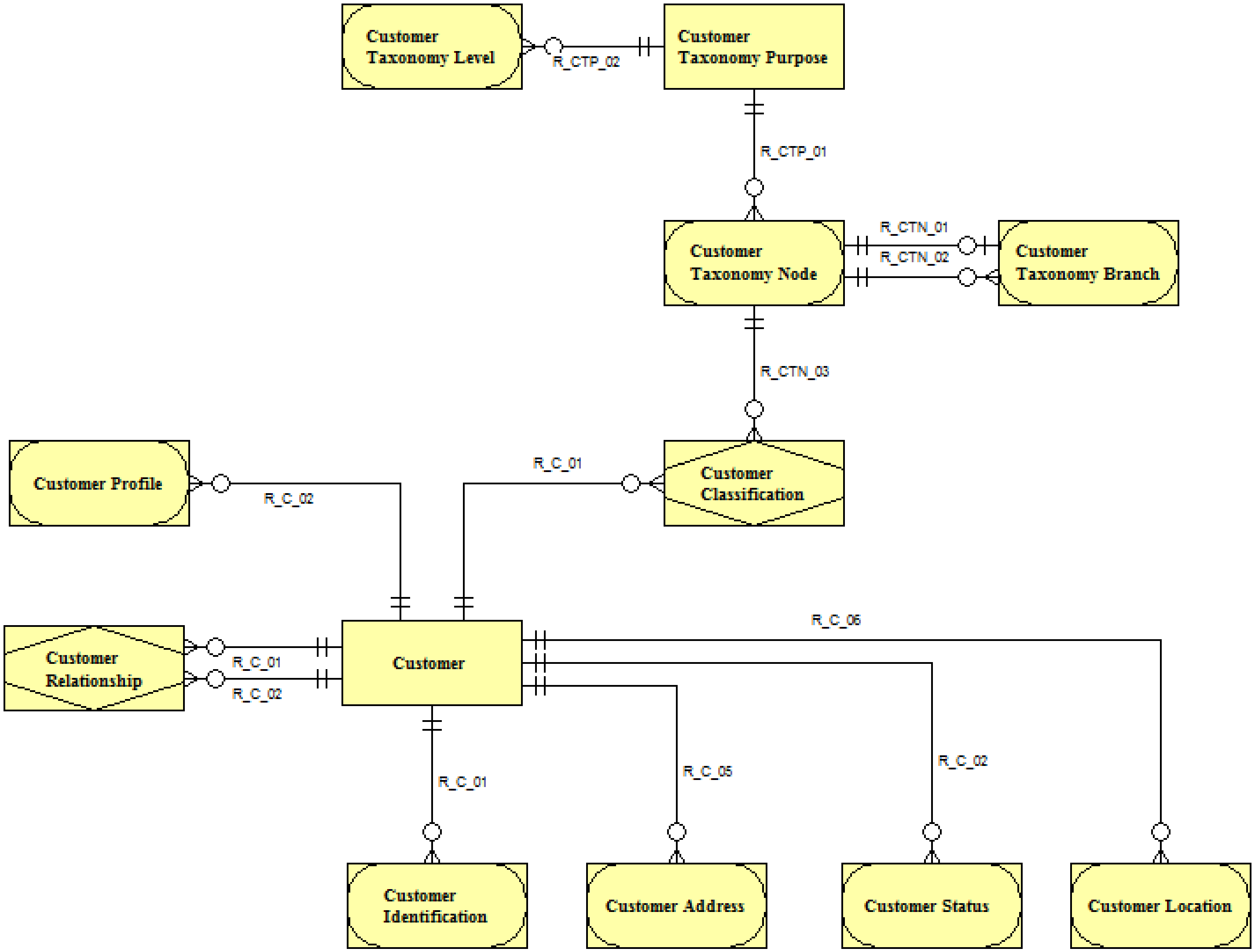
The worksheet shows a sample list of the relationship attributes. Other businesses may have more and different attributes and answers.

Entity Name	Attribute Name	Attribute Meaning	Logical Data Type	Default Value	Question 2 Relationship Target	Question 3 Single or Multi Valued	Entity Pattern Case
Customer	Customer Relationship	– the relationships that exist between different Customers	table key	zero	Customer	multi valued	6 Network Reflexive

The worksheet results would result in the customer relationship data model shown below.



Bringing the three example data models together would result in the consolidated diagram shown below.



6. REFERENCES

- Codd, Edgar, “A relational model for large shared databanks”, Communications of the ACM, Vol. 13, No. 6, Jun 1970
- Chen, P., "The Entity-Relationship Model-Toward a Unified View of Data"; ACM Transactions on Database Systems, Vol. 1, No. 1, pp. 9-36, Mar 1976
- Gotlieb, C.C. and Gotlieb, Leo R., Data Types and Structures, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978
- Mundy, J. and Thornthwaite, W. with Kimball, R, The Microsoft Data Warehouse Toolkit, Wiley Publishing, Inc., Indianapolis, Indiana, 2006
- Simsion, Graeme C. and Graham C. Witt, Data Modeling Essentials, 3rd Ed., Morgan Kaufman Publishers, San Francisco, CA, 2005
- Simsion, Graeme C., Data Modeling Theory and Practice, Technics Publications, LLC, Bradley Beach, New Jersey, 2007
- Visible Analyst Case Tool was used to create the data models.
-