EUROFILING Initiative

Cookbook

Data Point Modelling

Eurofiling Core Team

Table of content

[List of Figures II](#_Toc331596717)

[List of Tables II](#_Toc331596718)

[List of Abbreviations II](#_Toc331596719)

[1 Introduction 1](#_Toc331596720)

[2 Why data modelling is essential 2](#_Toc331596721)

[3 How can Data Points be modeled 4](#_Toc331596722)

[3.1 Goals 4](#_Toc331596723)

[3.2 Concept and Terms 4](#_Toc331596724)

[3.2.1 Measures 5](#_Toc331596725)

[3.2.2 Dimensions 5](#_Toc331596726)

[3.2.3 Domain 6](#_Toc331596727)

[3.2.4 Domain member 7](#_Toc331596728)

[3.3 Supervisory tables as starting point 9](#_Toc331596729)

[3.4 Normalization process 10](#_Toc331596730)

[3.4.1 Step 1 – Finding categories 11](#_Toc331596731)

[3.4.2 Step 2 – Making categories more abstract 12](#_Toc331596732)

[3.4.3 Step 3 – Normalizing 14](#_Toc331596733)

[3.4.4 Step 4 – Structuring 15](#_Toc331596734)

[3.5 Identification of hierarchies 17](#_Toc331596735)

[3.6 Further rules 17](#_Toc331596736)

[4 Visualisation of the results 18](#_Toc331596737)

[5 Summary 19](#_Toc331596738)

[List of References 20](#_Toc331596739)

[Contributors 20](#_Toc331596740)

List of Figures

Figure 1: Sample table 5

Figure 2: A cell in the cube corresponds to a cell in the table 6

Figure 3: Domain and Subdomain 6

Figure 4:Domain, Subdomain and Domain Member 7

Figure 5: Measure, Dimension, Domain, Subdomain and Domain Member 8

Figure 6: Sample table shape: original and with information made implicit 10

Figure 7: Different expressions for the same facts 11

List of Tables

Table 1: COREP sample hierarchy 9

List of Abbreviations

XML Extensible Markup Language

# Introduction

This document is oriented to banking specialists who are highly skilled in understanding supervisory reporting frameworks, and who are responsible for supporting IT experts to transfer the content of regulatory reporting and other supervisory purposes to IT systems.

To facilitate the communication between supervisory experts and IT experts, the Eurofiling initiative introduced the concept of Data Point Modelling in 2009. A Data Point Model (DPM) defines structures of data represented in supervisory tables that can be interpreted by IT applications. The DPM can generate data formats for (1) the reporting process or (2) to design multidimensional database structures for the analysis of supervisory data, i.e., in data warehouses.

This document aims to support business experts to create a DPM by introducing the concept of data point modelling as well as the underlying terms. Furthermore, it provides the basic knowledge about the procedure of normalising data. Normalisation is a key skill of Information Technology to transfer data from databases by reaching a greater consistency and by omitting redundancies, but these rules of normalisation are barely known by supervisory experts.

In addition to the normalisation of data, further rules must be taken into account for data point modelling. These rules enrich the DPM by providing knowledge about the relationship between financial concepts, especially for calculation and presentation purposes. These rules will be introduced by means of a “cookbook” for the creation of DPM for supervisory reporting.

In short, this document tries to explain how analysts get from a business table to a data structure that can be communicated and understood by computers. It is about the methodology that is called data point modelling, being the result of the process is a data point model.

# Why data modelling is essential

Up to now, the supervisory regulation is interpreted by experts from the supervisory authorities. They represent this knowledge in tables formated in the way that they like to view and analyse the data. These tables are often the basis for the reporting process. Each authority defines its own data format to collect the respective reporting data from its supervised entities. Often the data format is derived directly from the tables by giving the location of a financial piece of data, as represented by a combination of a row and column data point. In this case, the interpretation of the data in a report is only possible by referring back to the corresponding supervisory table.

However, such a data format requires a certain amount of maintenance:. (a) The column and row coordinates need to be stable and ascending; (b) Placeholders should ensure that new columns and rows can be added; (c) A presentation-based format also needs to deal with redundancies, which occur when items of detailed tables are repeated in summation tables: (d)Validation rules are needed to check that the same value is reported each time the item is repeated. The number of validation rules is high because a grouping based on ranges in tables is not possible when using row and column number combinations. A further disadvantage is the missing semantic meaning of the data reported which allows only an interpretation of the content in the file by viewing the according table and knowing the additional guidelines.

In 2005 IT representatives of different NSAs met to create a technical data format for the European reporting framework COREP. A normalisation[[1]](#footnote-2) of this data was initiated by IT experts to omit redundancies, reduce the maintenance effort and ease the extension of the format according to national discretions. During this normalisation process the reporting data was categorised and hierarchically structured. The data model ensured an efficient usage from an IT perspective but it lacks of a consistent structure from a supervisory expert point of view.

Over the course of time it was recognised that an unambiguous and consistent data model could increase the comprehensibility of the reporting data. Besides this a reuse of the model for analysis purposes could be achieved, especially in times of the financial crisis where detailed micro- and macro-prudential analyses are of high priority at every supervisory authority. It was obvious that IT experts are not able to achieve a DPM of good quality and that detailed supervisory knowledge is needed. First proposals of multidimensional data models were therefore created as joint venture of IT experts and supervisory professionals where IT knowledge about data modelling and normalisation was transferred and applied by supervisory experts to structure reporting data in a DPM. DPMs were published on the website of the Eurofiling initiative in 2009 for getting feedback and as proposal for upcoming reporting frameworks or new versions of existing ones.

# How can Data Points be modeled

## Goals

* comprehensibility
* basis for the identification of business data
* eases the implementation of interfaces
* explanation: multidimensional structures
	+ Star schema
	+ Data Warehouse
* omitting redundancies
* reaching consistency

## Concept and Terms

(editor: Katrin Heinze)

The concepts and terms when talking about data point modelling are inspired by the vocabulary used for multidimensional databases or data warehouses. Even so these terms of multidimensionality are IT technical terms they are also well known by financial experts i.e. in form of segmental breakdowns.

One case of application of multidimensionality is the visualisation of financial data in different views by using pivot tables. Pivot tables are able to aggregate the data and allow a compressed analysis. In the following the terms of multidimensionality are being introduced by offering an example which allows a common understanding between business and IT experts.

The sample table below illustrates how data can be divided in different categories or rather breakdowns. Here gross revenue is broken down by product group and country. Two implicit categories are not shown, the fiscal year and the company for which the results are presented.



Figure 1: Sample table

### Measures

*Measures* are properties which show the nature of the data. In general *measures* are numeric values and quantitatively measurable. They are therefore used on calculations and aggregations. Having only an amount of gross revenue with no additional information it could be interpreted as total current revenue but when referring to more detailed data “by” conditions are needed. Those "by" conditions are reflected in the example above "by" product group and "by" country.

### Dimensions

"By" conditions are called *dimensions*. *Dimensions* are views for which data should be stored and they can vary depending on the underlying framework. Some dimensions can more general ones like time or entity.

In the example above the country represents a list of possible countries but other dimensions also representing countries could be "country of origin" or "buying country".

**Dimension** describe the value and therefore restrict the possibilities of interpretation..

The table above shows the data in a two dimensional structure

Multidimensional data is often shown as three dimensional cube. The restriction on three dimensions is based on the limitations of the graphical representation.



Figure 2: A cell in the cube corresponds to a cell in the table



### Domain

A domain is a set of items having a specific coherence. This domain can be assigned to a dimension. A sub domain is a subset of items of a domain. A sub domain can also be assigned to a dimension, ie. European countries.



Figure 3: Domain and Subdomain

In the CEBS framework domains consists of all possible breakdown items like exposure classes or counterparties. If one table does not include all the possible items and using only a shortened list, a sub domain is created.

The elements of the Dimension (members) are used to describe the exact value and limit the scope for interpretation.

### Domain member

One item in the data set representing a domain is called domain member. In a CEBS taxonomy they are built up in a domain-member-ship.



Figure 4:Domain, Subdomain and Domain Member

Each combination of a measure with the according dimension - domain member combinations identifies a data set uniquely.

 Figure 5: Measure, Dimension, Domain, Subdomain and Domain Member

Hierarchies are sets of members of an explicit domain arranged in a hierarchical disposition. A node of a hierarchy can define basic arithmetical relationships (=, <= or >=) in relation to its child nodes. Child nodes can determine a sign (+ or -) that represents whether the member referenced should account positively or negatively to the arithmetical relationship defined in the parent node.

Hierarchies are a compact way of defining three ideas:

* Subsets of the members of a domain (sub-domains)
* Schematic arrangements of members of a domain for presentation purposes
* Basic arithmetic relationships

|  |
| --- |
| Approach for market risk |
| **=** |  | 1 | Total/NA |
| **=** | **+** | 2 | Standardised approaches |
| **=** | **+** | 3 | Normal treatment |
|  | **+** | 4 | Maturity-based approach |
|  | **+** | 4 | Duration-based approach |
| **=** | **+** | 4 | Other |
|  | **+** | 5 | Debt securities under the first category |
|  | **+** | 5 | Debt securities under the second category |
|  | **+** | 5 | Debt securities under the third category |
|  | **+** | 5 | Debt securities under the fourth category |
|  | **+** | 5 | Securitisation exposures subject to 1250% risk weighting or deduction and unrated liquidity facilities |
|  | **+** | 3 | Particular approach for CIUs |
|  | **+** | 3 | Additional requirements for options |

Table 1: COREP sample hierarchy

* explanation: parallel hierarchies

## Supervisory tables as starting point

* identified problems
	+ combination of data from different views
	+ presentation oriented
	+ number of dimensions
	+ granularity

## Normalization process

* what means normalization of data
* how can the data be structures

(editor: Ronald Hommes)

Below is a typical table that is formed by business users and is put before reporters to store the values that are needed to perform a task of analyzing figures. In this example the values have already been entered.

|  |
| --- |
| **Dimensional view on gross revenue.** |
|   | **Product group** |
| Books | News papers | Maga zines | Maps |
| **Country** | France | 500 | 452 | 124 | 35 |
| Spain | 852 | 634 | 236 | 85 |
| Austria | 632 | 234 | 963 | 45 |
| Belgium | 459 | 325 | 456 | 96 |

**More information made implicit**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | B | C | D |
| FR | 500 | 452 | 124 | 35 |
| ES | 852 | 634 | 236 | 85 |
| AT | 632 | 234 | 963 | 45 |
| BE | 459 | 325 | 456 | 96 |

Figure 6: Sample table shape: original and with information made implicit

The left table above represents information on product groups per country and from the title it is understood that the values are gross revenues. What the table doesn’t say, and what may be clear to the reporter or creator, is a host of data:

* Time; what timeframe are the values representing? A year, month, day?
* Unit of measure; what does 500 mean? Currencies? Units?
* Decimals; has there been rounding, how accurate is the number?
* Precision; are the values in thousands, millions?
* Reporting entity; are the values of a single company, a country, a branch?
* Definitions of the axes’ members; what is considered a book or a map?
* Status; are these figures final, approved by an accountant, budgeted?
* There may be more ...

From this list it is easily gathered that if reporter and creator are not on the same page huge discrepancies in the interpretation of the figures may occur.

The goal is to make this same table unambiguous. All the unanswered questions must be answered, there must be no room for interpretation by users on either end of the communication process of the figures of this table. All information must be made explicit.

In the process of listing all there is to know about this table, suddenly one may realize that the actual words on the axes’ are just one of many parameters that need to be explained in detail. In fact this same table could have drawn completely differently with expressing the exact same meaning, like this:

Reported by Oxford Book Ltd., fiscal year 2011

(A, FR) = Gross realized book revenue, in France 500.000 €

(A, ES) = Gross realized book revenue, in Spain 852.000 €

(A, AT) = Gross realized book revenue, in Austria 632.000 €

(A, BE) = Gross realized book revenue, in Belgium 459.000 €

(B, FR) = Gross realized news papers revenue, France 452.000 €

……………………………………………………….

Figure 7: Different expressions for the same facts

### Step 1 – Finding categories

By disseminating the information in the various tables by starting from the reported value, the computer will be enabled to generate all different layouts and still make the values accessible, and we’ll see that the table may be used in more closely related reports.

500 is:

* A monetary value
* In a currency
* In thousands
* Correct to two decimals
* Reported by a single commercial company
* Reported over a certain time span
* The realized figures as stated by the reporter
* It’s ‘gross’
* About a product group
* In a region (or country)

All these ‘words’, that form categories of information about the value 500, MUST be defined. For people some will so obvious that it is easily forgotten that a computer in itself is pretty stupid. Would NLG be allowed to be reported as a currency? It once was, but it is doubtful any user would expect a report nowadays containing this value.

Some of these categories are obviously formed with a closed list of values that are allowed. That list in itself may form the definition of the category. The category ‘a monetary value’ may be categorized by a mathematical expression like ‘all numbers 0 to 9 which may have a decimal point followed by two numbers 0 to 9’. Fortunately computers are excellent in understanding mathematics.

### Step 2 – Making categories more abstract

 (Step 2a? – Setting the reporting domain boundaries)

The next step is to make a higher abstraction level of the categories that are found representing the value ‘500’, if they represent the lowest level of granularity. This means that the category identified cannot be disseminated any further. This is arbitrary since the level of detail is determined by the need to have information reported. For instance, if Spain is the lowest level, there would be no need to define that Spain is made up by provinces or cities (or another geographical notion). On the other hand if an analyst would require the gross revenue in books in Madrid, Spain may be an indicator but the information required is unavailable.

If the level of granularity is determined based on a single table, most categories found will immediately be the lowest level. However, a single table is one in a series that form a report. Which in itself is just one in a reporting framework, which is just one in a line of business which is ... you get the idea. The trick here is to try to work out a boundary for what information is regarded ‘in the same domain’ to which the level of granularity can be set.

If a table on gross revenue is part of a report on sales figures in general the outcome on the level of granularity will be different than if the gross revenue is part of a report on profit and loss statements of a company. The fact that the domain boundary is arbitrary influences the level of abstract on the categories that are found on the value ‘500’.

An attempt to create the higher level of abstractness on the categories found:

* A monetary value: some mathematical data type
* In a currency: closed list of currencies allowed
* In thousands: closed list of ‘precision’ types that are allowed
* Correct to two decimals: closed list of the numbers of decimals of exactness
* Reported by a single commercial company: open list of all companies
* Reported over a certain time span: open list of all periods of time
* The realized figures as stated by the reporter: closed list of status of reported figures
* It’s ‘gross’: a list of at least ‘net’ and ‘gross’, maybe there are more
* About a product group: open list of all product groups reporting companies can invent
* In a region (or country): closed list of all countries as of a certain date

There is no doubt that many abstracts of categories will trigger people in claiming there are other levels of abstractness. This just means we need to agree, obviously it’s not always clear. To discuss a few:

Who determines the level of abstractness? Is it from the point of view of the reporter or from the point of view of the creator of the report? On the ‘Reported by a single company’ category there is a level of abstractness placed that indicates that it’s open. All companies may report and since that list is changing by the minute, it is impractical to make it a closed list. Unless of course the report is put before a closed set of members like a report from a central bank put before commercial banks. This list would be closed since the central bank is also the organisation that determines if a new commercial bank is allowed in the market. Again subjectivity enters and by setting the domain boundaries the verdict on the list being open or closed can be determined.

On the ‘In a region (or country)’ category a closed list is placed as the level of abstractness. In this case some extra wording is added ‘as of a certain date’. This introduces a new aspect that has to be taking into consideration. Some of the categories are changing across time. Once Yugoslavia was a country and Slovenia didn’t exist as a country. Nowadays (2012) that is different. So it is not just the value reported that has a time aspect on it, so do some of the categories. Again, if computers need to work with the data, it has to be made explicit what is meant.

***Definitions***

The category and its level of abstractness are considered ‘aspects’ on the value reported. The appropriate term for the value reported is ‘fact value’. The fact value and all of its aspects are called a ‘fact’.

### Step 3 – Normalizing

In simple wording the process of normalizing is about not having the same terminology multiple times in the reporting domain. There may be good reason to have two lists of countries for instance: one for representing where the reporting entity is active, and one for which region its products are being sold. Although there may overlapping countries for both purposes there can also be exclusive countries for either list. This just means there are two ‘lists’ or categories but the values in the list must not be duplicated. So there is only one ‘US’ which may or may not be part of one or both lists.

From this example one can deduce that the ‘members’ of a list are independent to a certain degree. They can participate in multiple lists. There in fact may be some kind if ‘super’ list in which all members are present but that list may even not be actively used in the reporting domain in a specific table since only its sibling lists are being used which use a subset each.

The reason for deduplication is simplicity, uniqueness and again unambigouty in the terms used. Everything is defined only once to satisfy computerized communication.

In theory every noun or combination of nouns could be a ‘member’ somewhere. Again the reporting domain determines if these nouns are eligible. If a single noun has multiple meanings it MUST be made unambiguous by combining it with other nouns. Eg. ‘Pig’ means a certain animal but some use it describe certain state officials. If both meanings are eligible in the reporting domain some noun has to be added to both. A word like this is called a homonym, a word with two meanings. Computers hate these words.

The opposite is a synonym two words that have the same meaning. Within the reporting domain either a single one is chosen, or the two being synonym are being made explicit by linking them somehow. Computers also dislike these words when their synonym is not being made explicit.

***Definitions***

A ‘list’ is called a dimension. An entry or member in a list is called a member.

### Step 4 – Structuring

Once the normalisation on all the aspects has taken place, there will be numerous lists that have members or that have been declared ‘open’ since the members are unpractical to list. There MUST be a dimension that contains all kinds of data types; basic types at the lowest level of granularity that look like nouns that identify the makeup of a fact. Types like monetary, date, string, etc. Some nouns are synonyms for these types, like value, text or period. This dimension is special. It contains the most elemental definition of a fact value.

You MUST have dimensions that appear in almost any table; dimensions that deal with time (both as fixed dates but also as words like current year, previous period etc.), the presentation of numerics (for dates like ccyy-mm-dd or dd-mm-yy, for monetary if a decimal point is used, or reporting takes place in millions, thousand etc.), the reporting entity in whatever structure it may manifest itself (as fiscal entities, legal entities, reporters own brake up in divisions or departments etc.)

What is left are the dimensions and their members that can be used on the data types. You may arrange the dimensions in logical groups which may even be hierarchical in nature. The purpose of structuring the dimensions is for other users than yourself to extend on the data collection. To be able to extend one would need a simple mechanism to find what is already present. A kind of index of what has been structured would help extenders to quickly find their way and prevent unnecessary additions to the model that has now been created.

***Definitions***

A measure is the name for a member in the special dimension that contains data types.

Numeric based measures MUST have a unit: the unit of measure. String based measures MUST have a language. This leaves dates, Booleans and binary data types.

***Rules:***

1. ***There shall be no hierarchy amongst members in a single dimension.*** It may be true that the world is divided in continents and these are divided in countries which are divided in regions which are divided in etc. But if your dimension to catch them was ‘countries’ than the regions are inappropriate. If your dimension was ‘regions’ than countries and regions are equivalent in nature. If there is overlap (some regions contain countries which are also listed separately) the normalization was insufficient.
2. ***There shall be no synonyms amongst members in the reporting domain.***
3. ***There shall be no redundancy in the reporting domain.***
4. ***Dimensions that only support a single member on a specific table are NOT a ‘default’ member.*** This member may be put on a Z-axis instead of the standard X and Y axes, but the dimension and its member need to be reported if the report is to be independent of other collections of data. Every bit of information MUST be made explicit in the report.
5. ***Nouns that are a member MUST NOT address the unit of measure.*** By having the member expressing the unit of measure, re-use of the member in other circumstances (other unit of measure) are made impossible.
6. ***Presentation is NOT an argument to structure dimensions and its members.***

## Identification of hierarchies

* identifying relations between financial concepts
* calculation rules as orientation

## Further rules

# Visualisation of the results

* visualization: only standard knowledge needed, only with standard tools

# Summary

List of References

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Data Point Modelling references at <http://www.eurofiling.info/dpm/index.shtml>

Contributors

2012-07-12 Katrin Heinze

2012-07-21 Allyson Ugarte

2012-08-01 Ignacio Boixo

1. Normalisation is the process of taking data from a problem and reducing it to a set of relations while ensuring data integrity and eliminating data redundancy

[http://db.grussell.org/section008.html]. [↑](#footnote-ref-2)