On the Emergence of Patterns for Spreadsheets Data Arrangements

Ricardo Teixeira

rd.teixeira@campus.fct.unl.pt

Vasco Amaral

vasco.amaral@fct.unl.pt

NOVA LINCS, DI, FCT, Universidade Nova de Lisboa, Portugal

Abstract. Spreadsheets are widely used both by individuals as well as large companies in a vast plethora of application domains. One of the reasons for this popularity is the general purpose flexibility spreadsheets offer to the end user. This flexibility favors the existence of multiple spreadsheet designs regarding the physical organization of the data presented by a spreadsheet. Nevertheless, to the best of our knowledge, little is still known about patterns of spreadsheet data arrangements. Works refer the emergence of commonalities and templates but it is hard to find a systematic study on the topic that presents us catalogues. It is known that spreadsheets are extremely error-prone. Therefore, to know the typical data arrangement patterns can be very useful insight on how to build mechanisms and strategies in order to prevent errors regarding spreadsheets specification and maintenance. The present work aims at present data arrangement patterns that emerged from our studies and direct observation of real-world spreadsheet samples from two large datasets, and, additionally, a formal representation of the patterns identified through the use of conceptual models.

Keywords: Spreadsheets, Data Arrangements, Patterns, Conceptual Model, UML

1 Introduction

Being the first "programmer in a box" to come along for technology users, spreadsheets are widely used both by individuals to cope with simple needs like tracking personal finances, training plans, to-do lists, supplier databases, or any purpose that requires input of data and/or performing calculations; as well as large companies as integrators of complex systems and as support for informing business decisions especially in areas like marketing, business development, sales, and finance. As result of this general purpose flexibility, a plenty of spreadsheet layout designs are possible towards the physical organization of the data composing a spreadsheet.

Works proposing spreadsheet models [1][2] already systematize common templates of table structures. Other works created a library containing common spreadsheet patterns [3] for later use of pattern matching algorithms in order to extract models from them. Other works implemented a header inference system for spreadsheets [4], describing the relation between the headers and their association with data.

However, these patterns are quite far from covering all existing kinds of spreadsheet's data arrangements and do not take in consideration the domains those patterns are generally applied.

Knowing more about the typical data arrangement patterns, in other words, what people usually want to model in a spreadsheet and what they usually expect to see in a spreadsheet, can be very useful insight in how to build mechanisms and strategies to specify and maintain less erroneous spreadsheets.

This work intends to take a step on extending the current perception of the emerged spreadsheet patterns regarding the data arrangements. For this purpose, two large repositories of spreadsheets used in spreadsheet studies were directly observed and analyzed, namely:

- The EUSES corpus [5] – published in 2005 and made available only to researchers, it is a dataset of over 4,500 spreadsheets gathered from the public world-wide-web;

- Enron corpus [6][7] – a recent large dataset containing around 15,000 industrial spreadsheets extracted from the Enron Corporation e-mail archive made public during the legal investigation concerning the company after it went bankrupt.

The analysis method consisted of manually selecting random spreadsheet samples from the datasets, until the patterns observed were becoming redundant. Due to the low diversity verified, only 80 spreadsheets representative of all of the spreadsheets existing in the datasets were selected and reunited. With them, a formal systemization of data arrangement patterns was made using the UML conceptual model, namely, class diagrams, which is one of the most proliferated conceptual models, having a high level of understanding.

The rest of the paper is organized as follows: in Section 2 we present the identified patterns, cataloging them and presenting related insights. Then, in Section 3 we present a metamodel of a spreadsheet concerning its data arrangement, and in Section 4 we conclude the paper.

2 Patterns

2.1 Table Structures

When thinking about spreadsheets we immediately conceive tabular forms constituted by a set of labels – usually called "headers" – associated with a set of values. Based on the spreadsheets observed, we can catalogue the common tables structures into three distinct groups which are defined by the table growth orientation and their purpose.

Vertical Tables.

The most linear table structure consists of a simple grown-vertically table, where there is a header in the first row; this structure is commonly associated with inventory, database (Fig. 1), or statistical data (Fig. 2). A header can represent a formula referring other row's entry values.

	A	B	C	D	E	F	G	H	I
1	**Title**	**First Name**	**Last Name**	**Country**	**Nationality**	**Age Group**	**Sex**	**Address_1**	**Address_2**
2	Ms.	Maxine P.	McClean	BARBADOS	Barbadian	35 - 44	Female	22 Oxnards Heights	St. James
3	Ms.	Jeanette	Bell	BARBADOS	Barbadian	45 - 54	Female	Women and Development Unit	School of Continuing Studies
4	Mr.	Anthony	Bovell	BARBADOS	Barbadian	45 - 54	Male	2nd Avenue Promenade Road	Kew Land
5	Dr.	Judith W.	Edwin	VIRGIN IS.	U.S.	> 55	Female	P.O. Box 306935	St. Thomas
6	Mr.	Raymond	Joseph	ST. LUCIA	St. Lucian	45 - 54	Male	Teacher Education Division	Sir Arthur Lewis Com. College
7	Ms.	Desiree V.	Edwards	ANTIGUA	Antiguan	35 - 44	Female	P.O. Box 1430	St. John's
8	Dr.	Edris L.	Bird	ANTIGUA	Antiguan	> 55	Female	P.O. Box 1810	St. John's
9	Ms.	Brenda C.	Carrott	ANTIGUA	Antiguan	25 - 34	Male	Lower Fort Road	St. John's
10	Ms.	Angela	Brice	ST. LUCIA	St. Lucia	> 55	Female	P.O. Box 4005	Castries
11	Ms.	Maureen	Lucas	BARBADOS	Barbadian	45 - 54	Female	Spooners Hill	St. Michael
12	Ms.	Martina	Augustin	ST. LUCIA	St. Lucian	45 - 54	Female	Sir Arthur Lewis Com. College	Morne Fortune
13	Ms.	Ruby	Yorke	ST.LUCIA	St. Lucian	> 55	Female	P.O. Box 1553	Castries
14	Ms.	Gem	Lynch	BARBADOS	Barbadian	45 - 54	Female	"Der-Land"	Upper Golf CLub Road
15	Ms.	Patricia E.	Linton	BARBADOS		45 - 54	Female	Lot # 2 Kirtons	St. Philip

Fig. 1. Vertical Table used as a database

Also, sometimes there is an additional bottom row that applies an aggregation function to some specific column, as we can see in Fig. 2.

	A	B	C	D
1	**Database Name**	**Searches**	**Full-text**	**PDF**
2	Academic Search Premier	118 964	101 189	32 644
3	American Heritage Children's Dictionary	2	1	0
4	Business Source Premier	26 768	21 656	7 353
5	Clinical Pharmacology	38	21	0
6	Funk & Wagnalls New World Encyclopedia	712	156	0
7	Health Source - Consumer Edition	5 297	980	125
8	Health Source: Nursing/Academic Edition	11 293	2 019	941
9	Image Collection	168	184	0
10	MAS Ultra - School Edition	4 180	1 829	59
11	MEDLINE	16 346	89	0
12	Military & Government Collection	3 325	609	44
13	Psychology and Behavioral Sciences Collection	32 346	7 536	3 567
14	Regional Business News	6 345	2 718	464
15	Religion and Philosophy Collection	5 827	682	186
16	Total	231 611	139 669	45 383

Fig. 2. Vertical Table used to display statistical data

Horizontal Single Entry Tables.

A second table structure is a table whose headers are disposed vertically, and in which there is only one entry. Typically, the purpose of this kind of tables is to display summary data, and usually an aggregation function is applied on the solo entry values.

In Fig. 3, a SUM function is used to calculate the "TOTAL INCOME" from the above entry values.

	A	B
1	The American Society of Hematology	
2	2002 Audited Financial Statement	
3		
4		
5	*INCOME*	
6		
7	Administrative	1 017 768
8	Annual Meeting	9 316 889
9	Awards	252 592
10	BLOOD Journal - Editorial	619 595
11	BLOOD Journal - Publishing	8 730 190
12	Clinical Research Training Institute	15 000
13	Education & Communications	111 759
14	Self Assessment Program	325 160
15		
16	*TOTAL INCOME:*	20 388 953

Fig. 3. Horizontal Single Entry Table example

Relationship Tables.

A third group of table structures are the relationship tables, consisting of tables that grow horizontally, with a highlighted header – the top one. The top header values are themselves headers, that is, without that header's entry value, the other header entry values are meaningless. Sometimes the top header label is omitted, being only displayed its values. Aggregation functions are also commonly used on this tables, both vertically (see row "8" in Fig. 4) and horizontally (see column "F" in Fig. 5).

This table structure pattern dominates spreadsheets used for financial modeling and analysis, with the top header usually representing calendar years (Fig. 4), year quarters, months, etc.

	A	B	C	D	E	F
1	Calendar Year	2002	2003	2004	2005	2006
2						
3	Volume / Day	100 000	100 000	100 000	100 000	100 000
4	Days/Year	365	365	365	365	151
5	Demand Charge/MMbtu	$ 0,356480	$ 0,356480	$ 0,299080	$ 0,299080	$ 0,299080
6	Gas Research Institute (GRI)	$ 0,001970	$ 0,001640	$ -	$ -	$ -
7						
8	Total Calendar Demand Charge	$ (13 083 425,00)	$ (13 071 380,00)	$ (10 916 420,00)	$ (10 916 420,00)	$ (4 516 108,00)

Fig. 5. Relationship Table using calendar years

2.2 Header Composition

In horizontal tables, it is usual to see headers composed by other headers. The main headers – the ones who are composed – typically represent categories, and the coupled ones are headers belonging to the category of the main header where they are attached.

Commonly, a main header's entry value consists of an aggregation function – usually SUM – applied to the coupled headers' entry values.

2004 FINANCIAL ANALYSIS					
	Q1	Q2	Q3	Q4	TOTAL
Expected number of purses sold:	500	600	700	800	2600
COSTS					
Cigar Boxes	**$ 250,00**	**$ 300,00**	**$ 350,00**	**$ 400,00**	**$ 1 300,00**
retaurants (1000 boxes for free)	$ -	$ -	$ -	$ -	
tobacco shops (1000 boxes for $1.00 each)	$ 250,00	$ 300,00	$ 350,00	$ 400,00	
Cigar Box Accessories ($3.00/box)	=(B4*3)	**$ 1 800,00**	**$ 2 100,00**	**$ 2 400,00**	**$ 7 800,00**
Resourses	**$ 13 850,00**	**$ 13 850,00**	**$ 13 850,00**	**$ 13 850,00**	**$ 55 400,00**
CEO/CIO ($25,000 each)	$ 12 500,00	$ 12 500,00	$ 12 500,00	$ 12 500,00	
Purse maker ($6.00/hour)	$ 1 350,00	$ 1 350,00	$ 1 350,00	$ 1 350,00	
Technology	**$ 704,00**	**$ 30,00**	**$ 30,00**	**$ 30,00**	**$ 794,00**
Web Site					
domain name	$ 35,00	$ -	$ -	$ -	
hosting	$ 30,00	$ 30,00	$ 30,00	$ 30,00	
digital camera	$ 300,00	$ -	$ -	$ -	
MS Access database	$ 339,00	$ -	$ -	$ -	
Macromedia Dreamweaver	$ 399,00	$ -	$ -	$ -	
Marketing	**$ 1 250,00**	**$ 1 250,00**	**$ 1 250,00**	**$ 1 250,00**	**$ 5 000,00**
Micellaneous Costs	**$ 1 000,00**	**$ 1 000,00**	**$ 1 000,00**	**$ 1 000,00**	**$ 4 000,00**
Total Costs	**$ 18 554,00**	**$ 18 230,00**	**$ 18 580,00**	**$ 18 930,00**	**$ (74 294,00)**
REVENUE ($60/purse)	**$ 30 000,00**	**$ 36 000,00**	**$ 42 000,00**	**$ 48 000,00**	**$ 156 000,00**
Total Revenue	**$ 30 000,00**	**$ 36 000,00**	**$ 42 000,00**	**$ 48 000,00**	**$ 156 000,00**
TOTAL PROFIT	**$ 11 446,00**	**$ 17 770,00**	**$ 23 420,00**	**$ 29 070,00**	**$ 81 706,00**

Fig. 6. Relationship Table with Coupling

In Fig. 6, we can see a relationship table composed by six main headers: "Expected number of purses sold:", "COSTS", "Total Costs", "REVENUE ($60/purse)", "Total Revenue" and "TOTAL PROFIT", with the last four ones consisting of formulas. The main header "COSTS" is composed by other six headers, with three of them – namely: "Cigar Boxes", "Recourses" and "Technology" – having attached headers of their own. It is also possible to verify that "COST" has no table entry values associated, functioning as a pure categorization label, meanwhile the lower level main headers, such as "Cigar Boxes", have entry values consisting of a SUM aggregation function applied to the headers' values they have attached.

2.3 Header Hierarchy

Similar to the composed headers, there are the hierarchically organized headers. Although in the header composition is express some sort of hierarchy, there are actually some major differences between the two header arrangements: in this type of header arrangement, the hierarchy is explicit, that is, the headers are not physically on the same

level; also, unlike composed headers, in this arrangement the top headers (the ones who have at least one header below in the hierarchy) do not have any values in the table associated to them; lastly, a header hierarchy appears in both vertical and horizontal table structures, although it is very uncommon to see it in a horizontal one.

	A	B	C	D	E	F	G	H	I
1	**IDAHO STATE CAPITOL ARTWORK AND DISPLAY INVENTORY**								
2									
3	**Category**	**Dimensions**			**Location**			**Description**	**Photo**
4		**Height**	**Width**	**Depth**	**Floor**	**Wing**	**Notes**		
5	portraits (10)	50"	35"		1	R	on walls between pillars; 8 on inside; 2 on outside, N side.	Territorial Governor portraits, there is a descriptive plaque under the William Wallace portrait on north side (plaque is 13"x10.5"). All belong to Historical Society.	none
6	sculpture	10'-6"	3'-1"	5'-6"	1	R	NE corner near stairway down	Miner statue, titled "The Patriot." Text on base reads, "Created by Kenneth Lonn, a Bunker Hill Mine Mechanic. This sculpture in steel is dedicated to the man and women of Idaho's mining industry. On loan to the State of Idaho By The Bunker Hill Company, Kellogg, Idaho A Subsidiary of Gulf Resources & Chemical Corporation." Height is approximate, to top of drill.	P002
7	display case	3'-3"	8'	5'-3.75"	1	R	NE corner near stairway down	Idaho State Capitol Plan, diorama under glass, on wood base.	P003
8	plaque	19"	18"		1	R	NE corner near stairway down	Detail about miner sculpture. Black plastic frame and cracked plexiglass cover.	P001
9	plaque	24"	20"		1	R	NE corner near stairway down	engraved bronze, "We Were Miners Then" by Gov. Phil Batt, next to miner detail plaque.	P001
10	plaques	18"	22"		1	R	outside of between-pillar wall, NW corner near Cap. Ed. Cntr.	Smaller engraved bronze plaque (3.75"x18.5") above says "In Memory of JFK." Larger plaque has Prayer of St. Francis of Assissi.	P004
11	picture	18.25"	22.5"		1	R	NW corner near Cap. Ed. Cntr.	Wood framed photo of USS Boise CL-47 ship. Back of item has very faded paper (unreadable) and stamp that says "Official United States Navy Photograph."	P005
12	picture	18.25"	22.5"		1	R	NW corner near Cap. Ed. Cntr.	Wood framed photo of USS Idaho BB-42 ship. Back has stamp that says "Official United States	P006
13	display case	37.5"	25.5"		1	R	wall on N side of entrance to W hallway	"Idaho Peace Officers Association" dark-stained wood display case with glass front. Contains 4 badges and several engraved name plates.	P007
14	plaque	33"	33"		1	R	outside of between-pillar wall, SW corner	Engraved bronze, "In memory of the deceased Idaho volunteers" who died in war of 1898-99 with Spain.	P008
15	plaque	32.75"	25"		1	R	outside of between-pillar wall, SW corner	Engraved bronze plaque with Gettysburg Address. Bottom dedication: "Presented by the Woman's Relief Corps Department of Idaho, to the State of Idaho, in honor of the grand army of the republic, September 11, 1928."	P009
16	plaque	32.75"	25"		1	R	outside of between-pillar wall, SE corner	Engraved bronze plaque, "Memorial Day Order" and description. Bottom dedication: "Presented by the Woman's Relief Corps Department of Idaho, to the State of Idaho, in honor of the grand army of the republic, September 11, 1928."	P010
17	sculpture	5'	3'	2'-3"	1	R	under stairwell in SE corner	Large stone on pedastal (dimensions 2'-5.5" tall x 2'-8" wide x 1'-8" deep). Plaque says "Dedicated 3/22/99, to be opened in year 2010." Another plaque says "Base dontated by CI."	P011
18	sculpture	6'	3'	3'	1	R	under stairwell in SE corner	bronze sundial on circular bronze base, some stones inside the "stalk" of the base, says "Anno 1974 on stalk	P012
19	display case	7'-3"	4'-1/2"	11.5"	1	R	SE corner by elevator	"Idaho Sheriffs Association," medium-stained wood case sitting on floor, glass front, contains sheriffs badges from all Idaho counties with large wood Idaho carving in center.	P013
20	plaque	32.5"	26.5"		1	R	SE corner by elevator	"American Mothers Inc. Idaho Mothers Hall of Fame" framed plaque	P014

Fig. 7. Vertical Table with a Header Hierarchy

	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2		**Quarterly Financial and Stock Information**											
3		Sony Corporation and Consolidated Subsidiaries											
4		Year ended March 31											
5		(Unaudited)											
6													
7			Yen in billions except per share amounts										
8			1st Quarter			2nd Quarter			3rd Quarter			4th Quarter	
9			2002	**2003**		2002	**2003**		2002	**2003**		2002	**2003**
10		Sales and operating revenue. . . .	\1 633,50	**\1 721,80**		\1 780,90	**\1 789,70**		\2 279,30	**\2 307,70**		\1 884,60	**\1 654,40**
11		Operating income (loss)	3,0	**51,9**		(3,4)	**50,5**		158,6	**199,5**		(23,6)	**(116,5)**
12		Income (loss) before											
13		income taxes	(14,3)	**116,6**		0,6	**48,8**		119,3	**201,9**		(12,8)	**(119,7)**
14		Income taxes	20,3	**53,6**		14,8	**(14,9)**		39,0	**65,5**		(8,9)	**(23,4)**
15		Income (loss) before cumulative											
16		effect of accounting changes . .	(36,1)	**57,2**		(13,2)	**44,1**		64,0	**125,4**		(5,5)	**(111,1)**
17		Net income (loss)	(30,1)	**57,2**		(13,2)	**44,1**		64,0	**125,4**		(5,5)	**(111,1)**
18		Per share data of common stock											
19		Income (loss) before cumulative											
20		effect of accounting changes											
21		-Basic	(\39,26)	**\62,23**		(\14,34)	**\47,89**		\69,72	**\136,19**		(\5,91)	**(\120,47)**
22		-Diluted	(39,26)	**57,90**		(14,34)	**44,70**		64,87	**126,05**		(5,91)	**(120,47)**
23		Net income (loss)											
24		-Basic	(32,75)	**62,23**		(14,34)	**47,89**		69,72	**136,19**		(5,91)	**(120,47)**
25		-Diluted	(32,75)	**57,90**		(14,34)	**44,70**		64,87	**126,05**		(5,91)	**(120,47)**

Fig. 8. Relationship Table with a Header Hierarchy

In Fig. 7 it is possible to see a vertical table with two header hierarchies (“Dimensions” and “Location”) which have a mere organizational purpose, with the intend to offer a clearer and focused table understating. However, header hierarchies can be use with a comparison purpose in mind. As we can see in Fig. 8, there is a hierarchy for each

header naming a year quarter (“1st Quarter”, “2nd Quarter”, “3rd Quarter” and “4th Quarter”) with all of them sharing the same semantic yet physically different sub-headers. Using this kind of arrangement obviates the need for multiple tables, whose physical separation makes it difficult to compare the analogous data from the distinct tables; or obviates the need for unique header labels – for instance, using “1st Quarter 2002”, “2nd Quarter 2002”, etc., that also complicates the data analysis.

2.4 Table Replication

In a spreadsheet, it is often observed the replication of table structures, only differing semantically in a certain aspect. In Fig. 9 we can see two structure replicas of a total of five replicas of a relationship table, only differing in the year in which the table data concerns. In this case, the replicas are distributed by different worksheets, however, the replication can also occur on a single worksheet as shown in the example in Fig.10, where to calculate the “INCOME” and the “EXPENSES” the same table structure can be used.

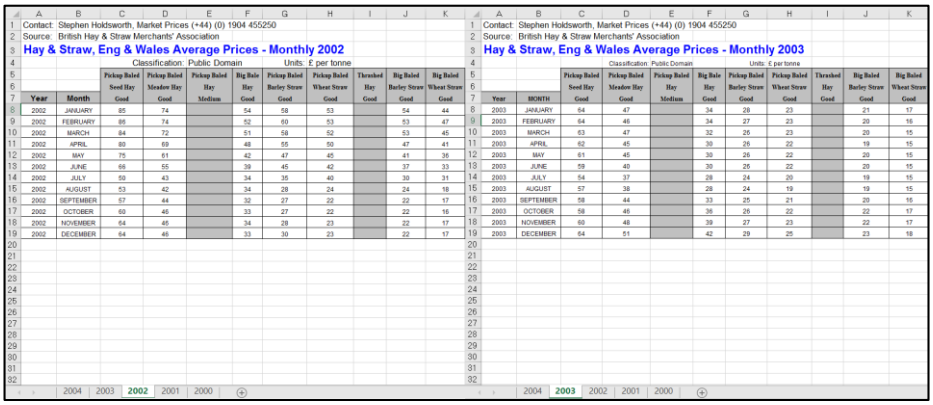

Fig. 9. Relationship Table replicated in different worksheets

The choice between the two replication options seem to depend on the table dimensions: larger table structures will naturally fit better in a spreadsheet on distinct worksheets (Fig. 9), while smaller ones can perfectly fit on the same worksheet (Fig. 10); and on the table purpose: if the spreadsheet analysis mainly relies on the comparison of the output data from the distinct replicas, it is convenient that the replicas stay physically close, which is the case of the example in Fig. 9 – besides the fact that the structures are quite small, the obvious object of analysis of the worksheet is the comparison between the “TOTAL INCOME” and the “TOTAL EXPENSES.

	A	B
1	The American Society of Hematology	
2	2002 Audited Financial Statement	
3		
4		
5	*INCOME*	
6		
7	Administrative	1 017 768
8	Annual Meeting	9 316 889
9	Awards	252 592
10	BLOOD Journal - Editorial	619 595
11	BLOOD Journal - Publishing	8 730 190
12	Clinical Research Training Institute	15 000
13	Education & Communications	111 759
14	Self Assessment Program	325 160
15		
16	*TOTAL INCOME:*	20 388 953
17		
18		
19	*EXPENSES*	
20	Administrative	1 313 117
21	Annual Meeting	4 719 488
22	Awards	1 633 459
23	Blood Journal - Editorial	1 252 340
24	Blood Journal - Publishing	5 492 425
25	Clinical Research Training Institute	5 216
26	Education & Communications	570 741
27	Self Assessment Program	378 892
28	CME	92 398
29	Committees	651 643
30	Development	200 774
31	International Members/Outreach	149 381
32	Membership Relations	512 101
33	Training Programs	193 705
34		
35	*TOTAL EXPENSES:*	17 165 680

Fig. 10. Horizontal Single Entry replicated in the same worksheet

3 A Metamodel for Spreadsheet Arrangement

The patterns identified in Section 2 can be formally systemized using and extending the UML conceptual model, specifically the UML class diagram metamodel. In Fig. 11, we present the metamodel in which spreadsheet elements – represented as entities – such as worksheets, tables, headers, etc., are an extension of the entity Class, and inherit some of its relations with other entities, namely, Association (with Aggregation and Composition specializations), Property and Usage.

The spreadsheet entities may have their own constants, for instance, the entity Worksheet have an integer constant named “order”. That constant indicates in which order the worksheet appears in the workbook, and so does the entity Table, but to indicate its placement in the worksheet relative to other tables. Additionally, Table has another constant named “Table Type” that specifies if the table grows vertically, horizontally, or if it is a relationship table.

Entities such as Table and Header can have Properties, which in the context of a class diagram are the commonly named Attributes. Those attributes specify child-headers, which can be further expanded to other headers, or be “leaf” headers.

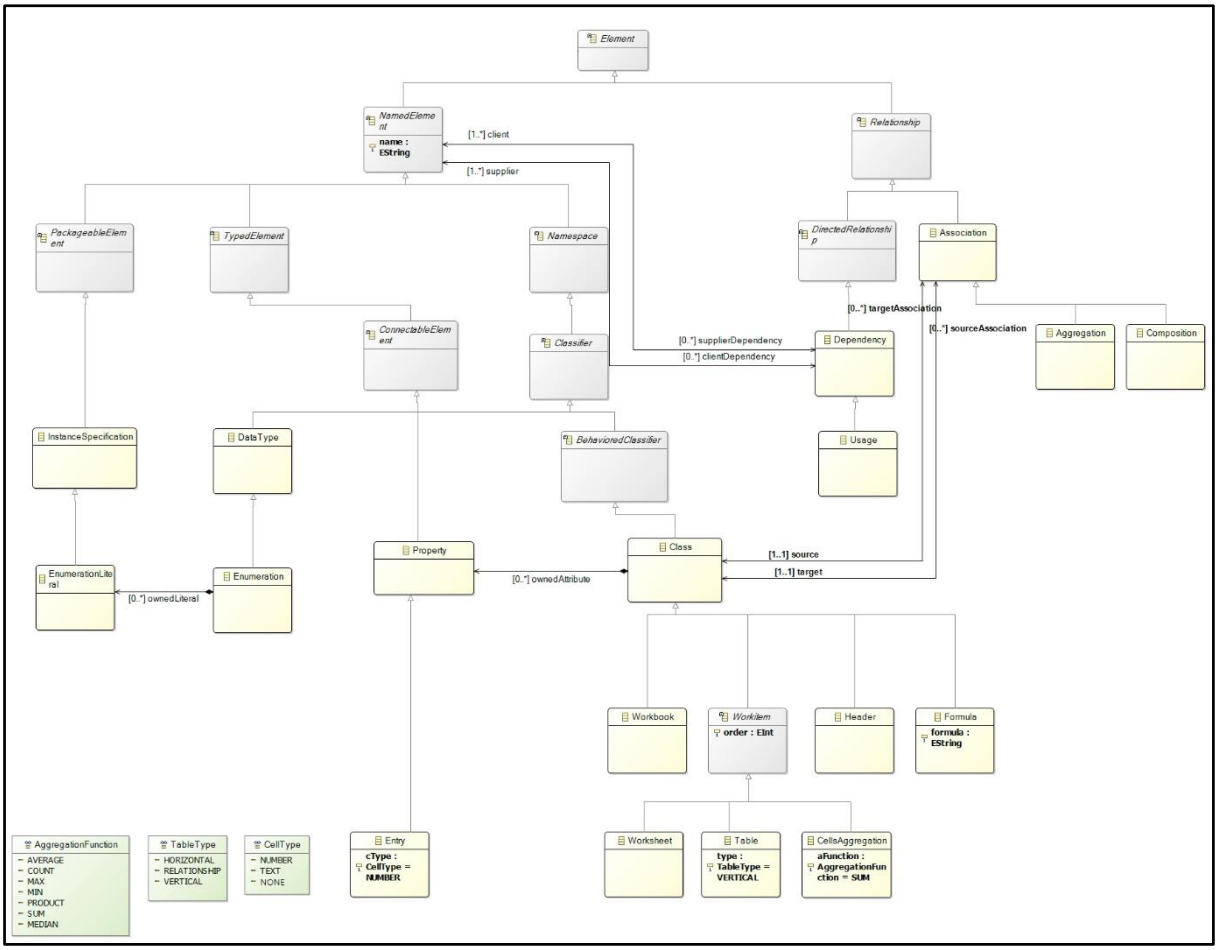

Fig. 11. Spreadsheet metamodel according to the detected patterns identified

With Association and its two extensions we can specify to which the spreadsheets entities connect and how this connection is done in terms of data arrangement. For instance, in Fig. 12 we can see a model (according to the metamodel) of the spreadsheet table shown in Fig. 7 of Section 2.3, where the header hierarchies are expressed through two aggregations. If there were no hierarchies, that is, all the headers placed on the same row, a composition would be used instead.

Using the entity Usage it is possible to specify usage dependencies among instances of the spreadsheet entities. For instance, as we see in Fig. 13 – a partial model of the table presented in Fig. 6 of Section 2.2 – there is an entity Formula to specify a formula associated to the attribute of the same name of the class to which this entity Formula is associated by a composition. This entity has a string constant to express the formula

text with the header reference between brackets. Moreover, there is expressed a dependency between the Formula entity and the corresponding header that is referenced, using Usage.

Furthermore, for a particular group of formulas, more specifically, the aggregation functions, there is a proper entity associated to the header of which attributes are input for the aggregation function specified in the entity CellsAggregation (see Fig. 14).

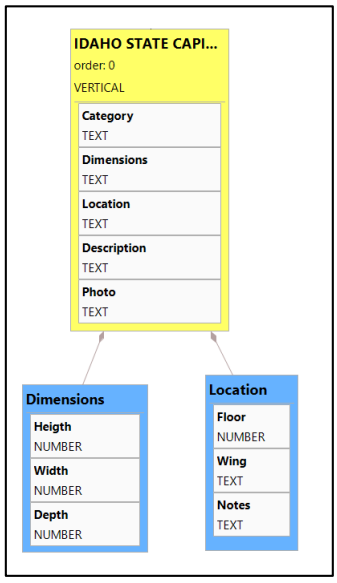

Fig. 12. Model representation of the table presented in Fig. 7 of Section 2.3

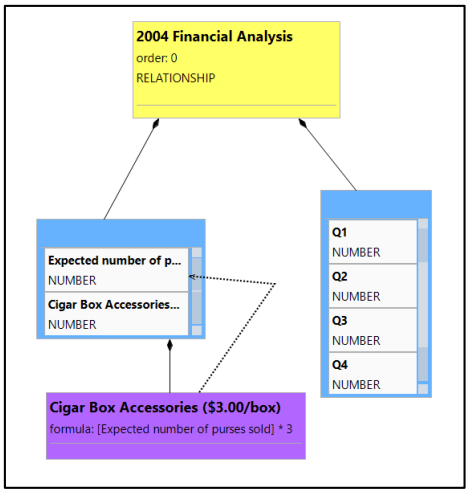

Fig. 13. Partial model representation of the table presented in Fig. 6 of Section 2.2

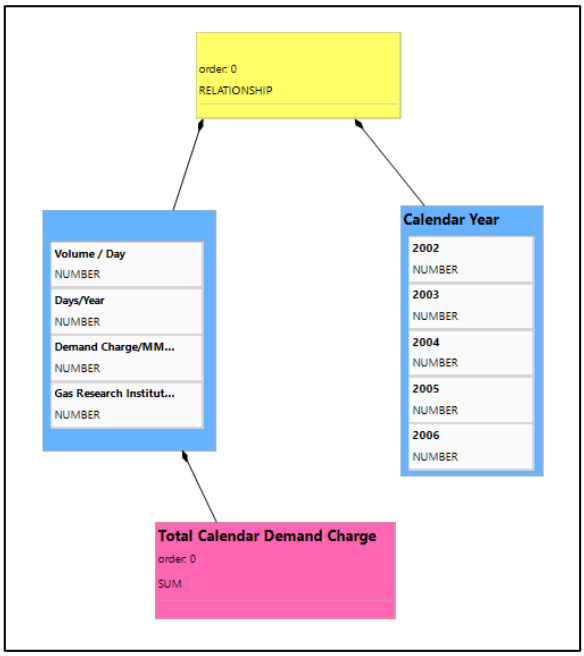

Fig. 14. Model representation of the table presented in Fig. 5 of Section 2.1

4 Conclusions

This paper presented a brief catalog of spreadsheet patterns regarding data arrangements layouts observed from two real-world spreadsheets datasets, extending and confirming the actual perceptions of the patterns in spreadsheets designs. Nevertheless, there is a major limitation on the approach taken, since neither of the datasets were fully covered, so it is possible that other existing patterns were not observed and, therefore, not registered. Moreover, this paper also presents a formalization of the identified patterns as a UML metamodel. This is an essential to design tools to build on top of the UML realm. In fact, the models we presented of the spreadsheets were created using a tool we implemented based on the metamodel. Conformance and other model-driven features are thus free to get.

Acknowledgements
This work has been partially supported by NOVA LINCS through the FCT project with reference UID/CEC/04516/2013.

5 References

1. Engels, G., and Erwig, M., *ClassSheets: Automatic generation of spreadsheet applications from object-oriented specifications.* In ASE '05: Proceedings of the 20th IEEE/ACM International Conference on Automated Software Engineering, pp. 124–133, New York, 2005

2. Cunha, J., Fernandes, J. P., Mendes, J., and Saraiva, J., *Embedding and Evolution of Spreadsheet Models in Spreadsheet Systems*. In Proceedings of the 2011 IEEE Symposium on Visual Languages and Human-Centric Computing, pp. 179–186, Pittsburgh, 2011

3. Hermans, F., Pinzger, M., and Deursen, A. van, *Atomatically extracting class diagrams from spreadsheets*. In Proc. of the 24th European Conference on Object-Oriented Programming, pp. 52-75, Berlin, 2010

4. Abraham, R., and Erwig, M., *Header and Unit Inference for Spreadsheets Through Spatial Analyses.* In Proceedings of the 2004 IEEE Symposium on Visual Languages and Human-Centric Computing, pp. 165-172, Rome, 2004.

5. Fisher, M., and Rothermel, G., *The EUSES spreadsheet corpus: A shared resource for supporting experimentation with spreadsheet dependability mechanisms*. In 1st Workshop on End-User Software Engineering, pp. 47–51, New York, 2005.

6. Hermans, F., and Murphy-Hill, E*., Enron's Spreadsheets and Related Emails: A Dataset and Analysis*, In ICSE' 15: 37th International Conference on Software Engineering, pp. 7-16, Florence, 2015

7. Jansen, B., *Enron versus EUSES: A comparison of two spreadsheet Corpora*. In SEMS'15: Second Workshop on Software Engineering Methods in Spreadsheets, pp. 41-47, Florence, 2015

Attachments

Attachment 1. EUSES' spreadsheet files

database
01_20_04.xls
consultants.xls
Database_excel95.xls
datadict.xls
dist_ed_courses_Jan2000.xls
document_de_reference#A828A.xls
EbscohostByDb2002-03.xls
epcdata2002.xls
FeatureList.xls
flip_usd5.XLS
FS_Upgrade_Plan_v3_111502.xls
FS_Upgrade_Proj_Mgmt_#A829F.xls
haymth.xls
haymth_old.xls
ps-cs-msc-new.xls
topconschedtemplate.xls

financial
02rise.xls
costfactors.xls
departmental_sales_e.xls
FinancialReport.xls
hist4q_e.xls
hist_e.xls
PersonalFinanceScope.xls
Prq403.xls
Q3_Final.xls
Q4_02.XLS
quaterly.xls
tab004.xls
treasurers_report_aud#A7EA4.xls
UF_Genetics_Financial#A7E51.xls
USFAthleticFinancialSummary.xls
W_SBT_financial.xls

grades
1A6EGrades.xls
262grades.xls
310Grades.xls
483_grades_web.xls

511Grades.xls

inventory
am-template-inventory.xls
capitol_art_inventory.xls
ColdStorage.xls
inventor.xls
Inventory%20Schedule%202004.xls
Inventory-Emergency_C#A84CC.xls
InventoryList.xls
NMfgInventory04.xls
nonstandby_inventory_#A8712.xls
Overview.xls
Software_inventory_sheet.xls
temp_videos0304.xls
TuftsGHGInventory.xls
VRSinventory01.xls
VRSinventory03.xls

Attachment 2. EURON's spreadsheet files

andrea_ring__4__BRLH Storage.xlsx
andrew_lewis__84__Notification Rpt 1200.xlsx
andy_zipper__109__Cost Allocation 02-21-01.xlsx
andy_zipper__112__mODEL 3 7 01 Base.xlsx
andy_zipper__115__DYNEGY-ICE VOL Jun1.xlsx
andy_zipper__266__Broker detail 5-29-01.xlsx
andy_zipper__290__AGA.xlsx
andy_zipper__342__COF Curves for Andy Zipper.xlsx
barry_tycholiz__870__EPNG BP Tariff Sheet.xlsx
benjamin_rogers__1003__NEPOOL-ZoneG Dailies.xlsx
benjamin_rogers__1024__TLR Analysis.xlsx
benjamin_rogers__1052__FPLE model.xlsx
benjamin_rogers__1058__newco development cash flow.xlsx
benjamin_rogers__1108__Wheatland O&M.xlsx
benjamin_rogers__1231__Comparison2.xlsx
benjamin_rogers__911__PJM Eastern Hub Pricing.xlsx
benjamin_rogers__936__PJM Model.xlsx
bill_williams_iii__1373__EOL 5-11.xlsx
bill_williams_iii__1395__EES September Daily.xlsx
chris_germany__2124__DecCohCHOICE-ENA.xlsx
chris_stokley__3947__NP15 DJ Charts.xlsx
darrell_schoolcraft__7827__imbalsumm0110.xlsx
larry_may__21636__ed052501.xlsx
louise_kitchen__22676__BGM 1024 ngpl.xlsx

phillip_m_love__30520__Paulacustomerlist.xlsx
stacey_white__39052__Summary Oct 15.xls
steven_p_south__39352__04-23-01 Earnings 2 of 2.xlsx
vladi_pimenov__41075__VLADI-GASDAILY-CURVEFETCH.xlsx