

### Reverse Engineering of Data Models from Legacy Spreadsheets-Based Systems: An Industrial Case Study

discussion paper



Domenico Amalfitano Anna Rita Fasolino Valerio Maggio Porfirio Tramontana Vincenzo De Simone

# Spreadsheet Based Information System Issues

- Spreadsheets are designed only for computing purposes and commercial applications but ...
- ... very often they are used as Information Systems
  - Very difficult to maintain
    - High rate of duplicated data between different sheets and files
- The first and more critical step of a migration process is the Data Reengineering

# Case Study

- An automotive company collects the specification of the tests executed on the vehicles in form of Test Patterns
  - Test Patterns are implemented in Excel files following a common template
- We have 30,615 different Excel files with 2,700 data cells on average
  - There is a high rate of replication data
    - 50% of data cells recurred more than 100 times.
- Excel Test Patterns represent the input of an automatic test generation process

Time (s)   Every	N-1 5, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,	30-1 5,1 30-1 5,1 30-1 5,1 30-1 5,1	Stop	Start	Dusy Cycle	Freq. Otcl		Signal	sage	DE LOS	er i e	Inset Mag		Delete Step	Ireari Step
FIRE	N-1 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,	3/4 5/1 3/4 5/1 3/4 5/1								nel	7.	Description	1	Ingrat Manne	Time (s) / Evere
STATUS_B_CPL_DMFALSTS == 1 AMD 21 == 12    IPGETCAMSRADAUAS_FERRY   IPASSADIN   ATVOCALA SCHEPMATA   1 NA. BLCAM 1994_A_FUS.   Average full control for the 1   24-1   1   1   1   1   1   1   1   1   1	38-1 5; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1;	X-1 5,1							OD.	PRODUCE M					
R_CAM   SPE_A_FUSE   Amongsth action of Photology   SPE	N-1 5, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,	341 51							OB.	PRODEL M					
B_CAM   SPE_A_FULS   developed action of Photology   SPE_A	N-1 1; 1; 1; 1; 1; 1; 1; 1; 0; geals No PreqDi		_						_A_PLUS	B_CAN TE	1 W	ATWOALLA SCHERMATIF	[VF163M03]	TP_GETCAMERADA/JABUFFER()	(STATUS_B_BCMEMSFALSES 1 AMD 21 12)
MEY	FreqDi	34 51							LA_PUUS	BLCAN TE					
MEY	FreqDi No No								DEL.	PRODEL M					
NEY   PREV_DRI	No.						ÚT	PRIST_PAMOE_VALUE_OUT	DEL.	PRODEL M	$\overline{}$				
E.CAM 1997_A.P.U.S   develope Automotive Dates     PRODUCT NOCE	No		1	0	[70.0]	11000			DEL.	PRODEL M					
PRODE_   NOCE_   CAMERA_RANGE_OUT   PRODE_STATUS_D_ECM_DANGE_CO    PRODE_NOCE_   REST_RANGE_VALUE_OUT     PRODE_NOCE_   REST_RANGE_VA	No. Ugusie						Sur	AverageFueConsuptionAPlu	LA_PLUS	BLCAN TO	l W	State chiase	(- KEY_DN)	KEY	MEY
PRODUCT   NOCE   POST PRINCE CAT	Uguale						idData	AverageFueConsAPlusValid	LA_PLUS	B_CAN TO					
SETWIND   WWP_TAME-PAR_TAME-FULLETSTATO DEFINITIO COM   1								CAMERA_RANGE_OUT							
MR_TAMP FAR_TAMP FALLES STATO DEPORTO COM   1   NA.   B.C.AM 1994_A. F.U.S.   desempt factor control of the company of the c							úT		OD.	PRODEL M					
P. CAM   THE A. P. U.S.   Account of the Video     PERCE,   VICEO,   CAMPILL PARKE, CUI     PERCE,   VICEO,   PERCE,   VICEO,   PERCE,   VICEO,	16								OBL	PRODEL M		1. /			
	144						No	AverageFuelCorespitorAPlu			1 IW	&ESTATO DEFINITO COM	MAR_TIME:PAR_TANKFULL\$	SETWARD	
PRODEL MODEL PROST_MAKEE_VALUE_OL PRODEL_XEV OIL	No.						loData								
PRODEL MODEL PROST_MAKEE_VALUE_OL PRODEL_XEV OIL	140								DEL.	PRODEL M					
	Ugush						MODEL_REVOID	PRIST_PANCE_VALUE_OU	DEL.	PRODEL M	$\neg$				
MODEL MODEL PARKE VALUE ALOUT	No								OEL.	PRODEL M					
SETWAR) [IVAP_PROVEPAR_PROVED ESTATO DEFINITO COM 1 NA G.CAN TRP_A_PUS Average/us/Consiston/APTRO_SIGNA_TRANSPORCE_ENTER_ENABLE_NEG	Ugusè					2016					l N	É STATO DEPINTO COM	(VAR_PROXIPAR_PROXI)	SETVARIO	
E_CAN   1987 A_FULS   Average Author/Modificate   FORCE   NOCE   CANEER, A FULS   Average Author/Modificate   FORCE   FORCE   FORCE   Average Author/Modificate   FORCE   FORCE   FORCE   Average Author/Modificate   FORCE   FORCE   FORCE   FORCE   Average Author/Modificate   FORCE   FORCE   FORCE   FORCE   Average Author/Modificate   FORCE   FORCE   FORCE   FORCE   Average Author/Modificate   F	No.						idData	AverageFueConsAPlusValid	A_PLUS	B_CAN TR	- [				

### Data Model Reverse Engineering

 Data Model Reverse Engineering is the first step of a more general migration process towards a Web MVC architecture

 An heuristic based approach to infer the Data Model was proposed.

- A set of 26 heuristics were considered.
  - 11 heuristics derived from the literature and were adapted to work in this specific context.

## Data Model Reverse Engineering

 Heuristics can be grouped in two main classes:

- Structure based rules (SBRs)
- Information based rules (IBRs)

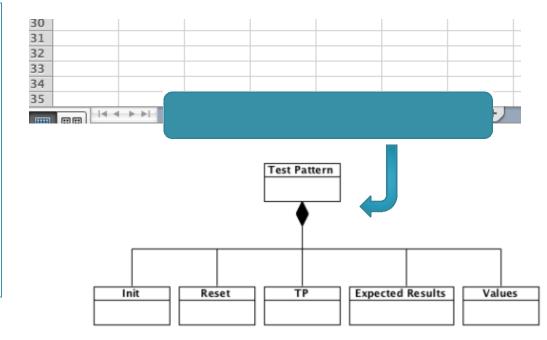
## Structure based rules (SBRs)

- SBRs analyze the structure and the properties of spreadsheets and their components, such as sheets, cells, cell headers, etc.
  - Used to abstract the set of candidate classes and their relationships;
  - Applied to a single Excel File.

## Example of SBR

#### Rule:

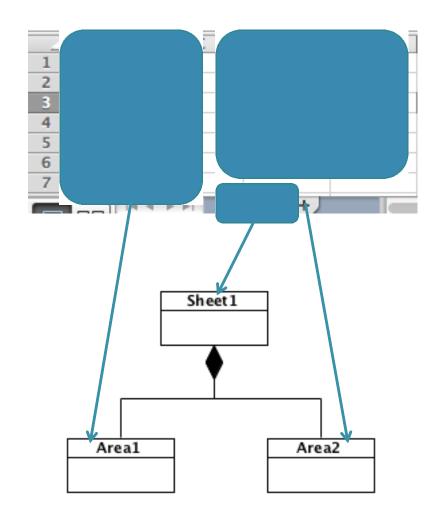
If the spreadsheet contains more than one sheet, then it is possible to associate the spreadsheet to a class C and each component sheet to a distinct class S<sub>i</sub>, where C has a UML composition relationship with each S<sub>i</sub>.



# Example of SBR

#### Rule:

If a sheet S contains sets of consecutive non-empty cells (hereafter non-empty cell area) that are well delimited from each other by means of empty cells, then it is possible to associate each non-empty cell area to a single class  $C_i$  and the sheet S to a candidate class  $C_S$ , where S has a UML composition relationship with each  $C_i$ .



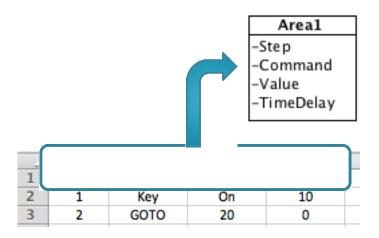
# Information based rules (IBRs)

- IBRs analyze the informative content of the cells by looking for repeated data, synonyms, and cells containing well-defined data structures such as array strings, integer matrixes, etc.
  - Used to infer the attributes of classes, the relationships between classes and their cardinalities.
  - Applied to all the Excel Files

# Example of IBR

#### Rule:

If the header cells of the columns that discriminated the extraction of a given class A assume the same textual content in all the spreadsheets, then these values may be considered attributes of that class.



# Process Execution and Results

- Selected groups of rules were iteratively applied to the spreadsheets.
- Sets of candidate classes and relationships were automatically proposed.
- The data model made by 18 classes, 27 relationships, and 95 attributes was reconstructed at the end of the process.
- Candidates were submitted to domain experts who chose to accept, to refine or to reject them.
  - Experts accepted 75% of candidates inferred by means of SBRs and 33% of candidates inferred by IBRs