A Toolkit for Scalable Spreadsheet Visualization

Markus Clermont
SQRL
University of Limerick

Contents

Software Quality Research Laboratory
University of Limerick

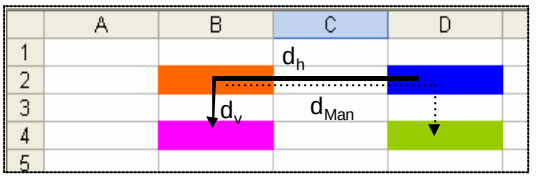
Motivation

- Many methods for spreadsheet-
  - design
  - implementation
  - testing

  have been suggested
- Not widely used, because
  - The way spreadsheets can be used is restricted
  - Involves software engineering terminology
- Visual Auditing can be done straight forward and helps
  - To find different kinds of errors
  - To comprehend a spreadsheet

Semantic Classes

- Idea:
  - Large spreadsheets are usually created by copy, paste & modify operations.
  - Aims to identify parts that are copies of the same original
  - Does not intend to change the way spreadsheets are created
- Similar Semantic Units form a Semantic Class
  - A Semantic Unit is a set of adjacent cells
  - Parameters d_h, d_v, d_{Man}



Semantic Classes

- Similar Semantic Units form a Semantic Class
  - Notion of Similarity:
    - Same geometrical shape
    - Cells on the same relative position in similar semantic units have similar formulas
    - Similar formulas → Logical Areas
  - Formulas can be
    - Copy-,
    - Logical, or
    - Structural equivalent

Semantic Classes

	A	B	C	D	E
1		Revenue	Expenses	Difference	Total Earnings
2	1	150	75	75	75
3	2	324	100	224	299
4	3	220	90	130	429
5	4	800	300	500	929
6	5	400	120	280	1209

	A	B	C	D	E
1		Revenue	Expenses	Difference	Total Earnings
2	1	150	75	=B2-C2	=D2
3	=A2+1	324	100	=B3-C3	=E2+D3
4	=A3+1	220	90	=B4-C4	=E3+D4
5	=A4+1	800	300	=B5-C5	=E4+D5
6	=A5+1	400	120	=B6-C6	=E5+D6

Semantic Classes

	A	B	C	D	E
1		Revenue	Expenses	Difference	Total Earnings
2	1	150	75	75	75
3	2	324	100	224	299
4	3	220	90	130	429
5	4	800	300	500	929
6	5	400	120	280	1209
7					

	A	B	C	D	E
1		Revenue	Expenses	Difference	Total Earnings
2	1	150	75	=B2-C2	=D2
3	=A2+1	324	100	=B3-C3	=E2+D3
4	=A3+1	220	90	=B4-C4	=E3+D4
5	=A4+1	800	300	=B5-C5	=E4+D5
6	=A5+1	400	120	=B6-C6	=E5+D6
7					

Semantic Classes

	A	B	C	D	E
1		Revenue	Expenses	Difference	Total Earnings
2	1	150	75	75	75
3	2	324	100	224	299
4	3	220	90	130	429
5	4	800	300	500	929
6	5	400	120	280	1209
7					

	A	B	C	D	E
1		Revenue	Expenses	Difference	Total Earnings
2	1	150	75	=B2-C2	=D2
3	=A2+1	324	100	=B3-C3	=E2+D3
4	=A3+1	220	90	=B4-C4	=E3+D4
5	=A4+1	800	300	=B5-C5	=E4+D5
6	=A5+1	400	120	=B6-C6	=E5+D6
7					

Software Quality Research Laboratory
University of Limerick

Semantic Classes

	A	B	C	D	E
1		Revenue	Expenses	Difference	Total Earnings
2	1	150	75	75	75
3	2	324	100	224	299
4	3	220	90	130	429
5	4	800	300	500	929
6	5	400	120	280	1209
7					

	A	B	C	D	E
1		Revenue	Expenses	Difference	Total Earnings
2	1	150	75	=B2-C2	=D2
3	=A2+1	324	100	=B3-C3	=E2+D3
4	=A3+1	220	90	=B4-C4	=E3+D4
5	=A4+1	800	300	=B5-C5	=E4+D5
6	=A5+1	400	120	=B6-C6	=E5+D6
7					

Semantic Classes

Auditing Strategy:

1. Identify Semantic Units & Classes
2. For each Semantic Class, test one unit in detail
3. Check for geometrical patterns
4. Find irregularities / ruptures
5. Mark them as *hot-spot* for further checking

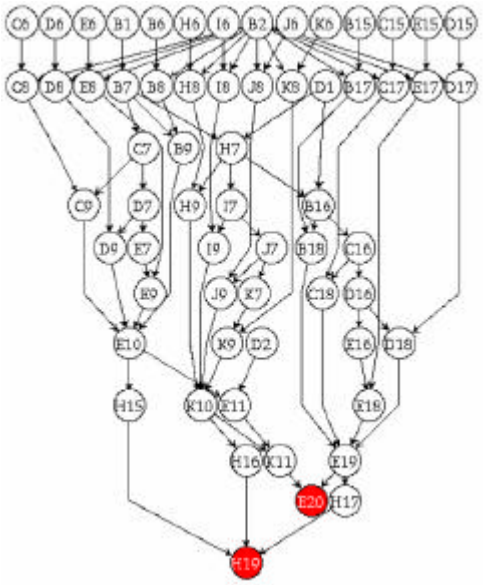
Data Modules

- Partition the DDG of a spreadsheet in a way, that
  - Each data module has one result cell
  - Other cells reference only the result cell of the data module
- Bottom-up construction, starting at the result cells
- Identification of result cells
  - Sink nodes of the pre-processed DDG
- A cell that influences only one result is added to its data-module
- Intermediate results are cells that influence more than one final result
  - Intermediate results are seeds for their own data module

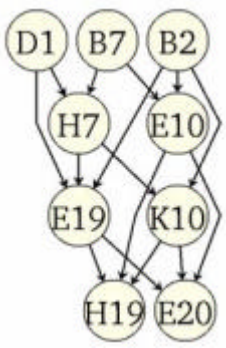
Data Modules

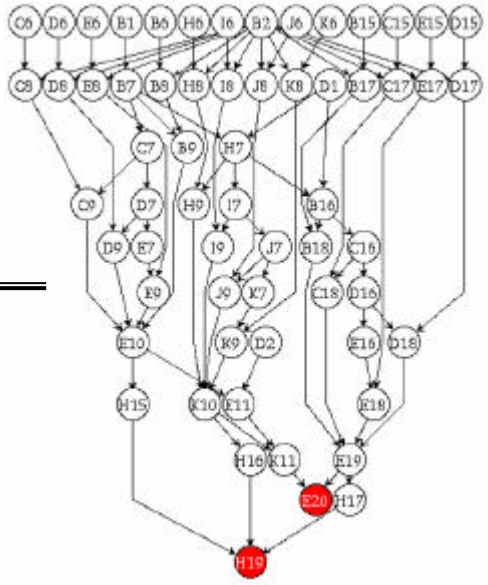
	A	B	C	D	E	F	G	H	I	J	K
1	Fixed Cost:	3000	Inflation:	0,02							
2	Earning/Unit	20	Cash:	1000							
3											
4	**Year**	1999					**Year**	2000			
5		1.Quarter	2. Quarter	3. Quarter	4. Quarter			1.Quarter	2. Quarter	3. Quarter	4. Quarter
6	**Sales**	400	250	100	450		**Sales**	250	300	50	300
7	**Fixed Cost**	3000	3000	3000	3000		**Fixed Cost**	3060	3060	3060	3060
8	**Earning**	8000	5000	2000	9000		**Earning**	5000	6000	1000	6000
9	**Result**	5000	2000	-1000	6000		**Result**	1940	2940	-2060	2940
10	**Result 1999:**				12000		**Result 2000:**				5760
11	**Cash 1999:**				13000		**Cash 2000:**				18760
12											
13	**Year**	2001									
14		1.Quarter	2. Quarter	3. Quarter	4. Quarter		Overview:				
15	**Sales**	300	200	370	100		Result 1999	12000			
16	**Fixed Cost**	3121,2	3121,2	3121,2	3121,2		Result 2000	5760			
17	**Earning**	6000	4000	7400	2000		Result 2001	6915,2			
18	**Result**	2878,8	878,8	4278,8	-1121,2						
19	**Result 2001:**				6915,2		Overall	24675,2			
20	**Cash 2001:**				25675,2						

Data Modules

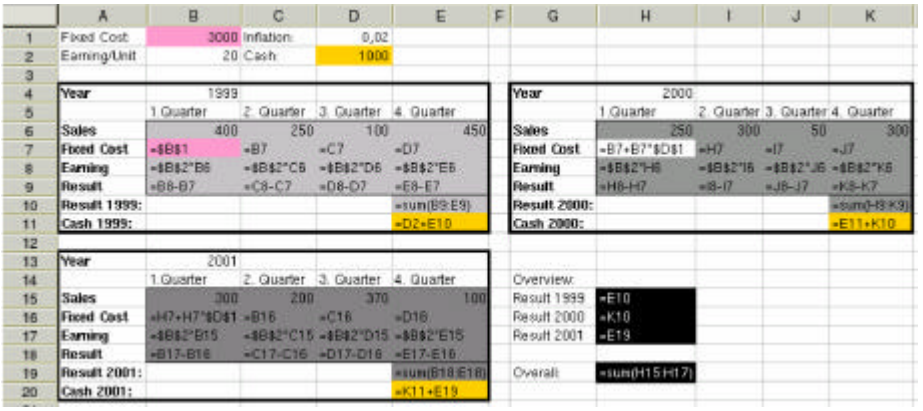


Data Modules





Data Modules



	A	B	C	D	E	F	G	H	I	J	K
1	Fixed Cost:	3000	Inflation:	0,02							
2	Earning/Unit	20	Cash:	1000							
3											
4	Year	1999					Year	2000			
5		1.Quarter	2. Quarter	3. Quarter	4. Quarter			1.Quarter	2. Quarter	3. Quarter	4. Quarter
6	Sales	400	250	100	450		Sales	250	300	50	300
7	Fixed Cost	=$B$1	=B7	=C7	=D7		Fixed Cost	=B7+B7*$D$1	=H7	=I7	=J7
8	Earning	=$B$2*B6	=$B$2*C6	=$B$2*D6	=$B$2*E6		Earning	=$B$2*H6	=$B$2*I6	=$B$2*J6	=$B$2*K6
9	Result	=B8-B7	=C8-C7	=D8-D7	=E8-E7		Result	=H8-H7	=I8-I7	=J8-J7	=K8-K7
10	Result 1999:				=sum(B9:E9)		Result 2000:				=sum(H9:K9)
11	Cash 1999:				=D2+E10		Cash 2000:				=E11+K10
12											
13	Year	2001									
14		1.Quarter	2. Quarter	3. Quarter	4. Quarter		Overview:				
15	Sales	300	200	370	100		Result 1999	=E10			
16	Fixed Cost	=H7+H7*$D$1	=B16	=C16	=D16		Result 2000	=K10			
17	Earning	=$B$2*B15	=$B$2*C15	=$B$2*D15	=$B$2*E15		Result 2001	=E19			
18	Result	=B17-B16	=C17-C16	=D17-D16	=E17-E16						
19	Result 2001:				=sum(B18:E18)		Overall:	=sum(H15:H17)			
20	Cash 2001:				=K11+E19						

Data Modules

Auditing Strategy

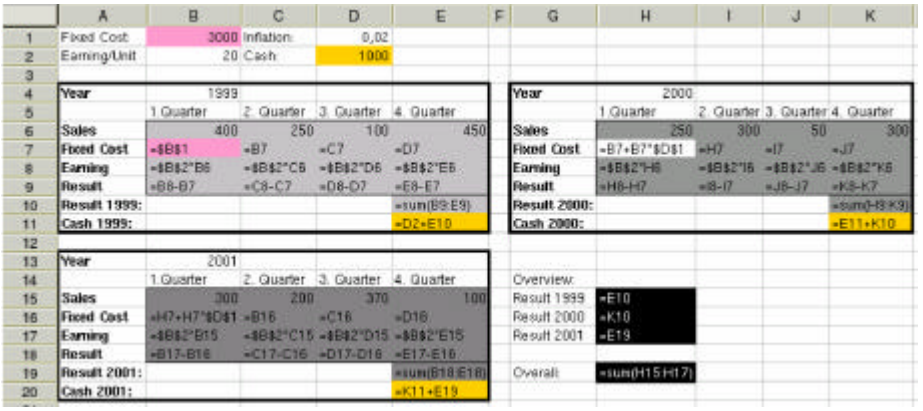
1. Identify Data Modules
2. Find superfluous data modules
   - Indication of misreferences
3. Check, whether all expected data modules are part of the visualization
4. Compare layout blocks with data modules

Data Modules

Fault Tracing

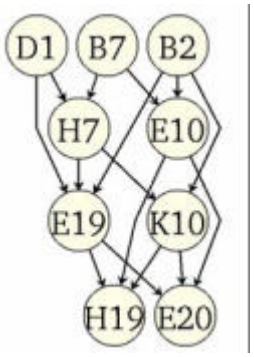
- identify modules that are referenced by a faulty cell
- check the result
  - If the result is correct, the module is correct and the error must be in another module
  - Else the error is either in the module, or in a referenced module.

Data Modules

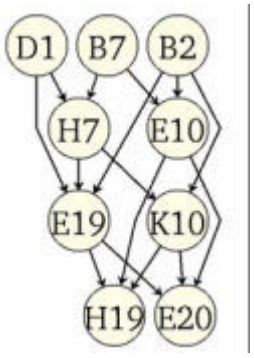


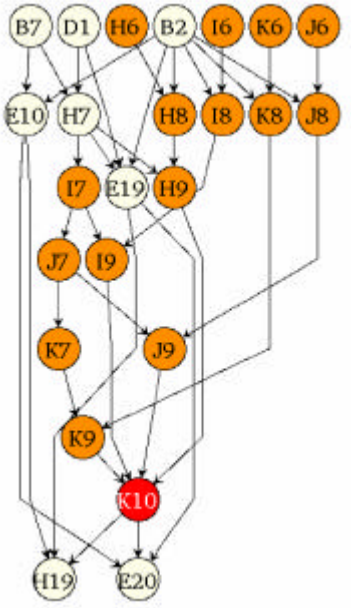
	A	B	C	D	E	F	G	H	I	J	K
1	Fixed Cost:	3000	Inflation:		0,02						
2	Earning/Unit	20	Cash:		1000						
3											
4	Year	1999					Year	2000			
5		1.Quarter	2. Quarter	3. Quarter	4. Quarter			1.Quarter	2. Quarter	3. Quarter	4. Quarter
6	Sales	400	250	100	450		Sales	250	300	50	300
7	Fixed Cost	=$B$1	=B7	=C7	=D7		Fixed Cost	=B7+B7*$D$1	=H7	=I7	=J7
8	Earning	=$B$2*B6	=$B$2*C6	=$B$2*D6	=$B$2*E6		Earning	=$B$2*H6	=$B$2*I6	=$B$2*J6	=$B$2*K6
9	Result	=B8-B7	=C8-C7	=D8-D7	=E8-E7		Result	=H8-H7	=I8-I7	=J8-J7	=K8-K7
10	Result 1999:				=sum(B9:E9)		Result 2000:				=sum(H9:K9)
11	Cash 1999:				=D2+E10		Cash 2000:				=E11+K10
12											
13	Year	2001									
14		1.Quarter	2. Quarter	3. Quarter	4. Quarter		Overview:				
15	Sales	300	200	370	100		Result 1999	=E10			
16	Fixed Cost	=H7+H7*$D$1	=B16	=C16	=D16		Result 2000	=K10			
17	Earning	=$B$2*B15	=$B$2*C15	=$B$2*D15	=$B$2*E15		Result 2001	=E19			
18	Result	=B17-B16	=C17-C16	=D17-D16	=E17-E16						
19	Result 2001:				=sum(B18:E18)		Overall:	=sum(H15:H17)			
20	Cash 2001:				=K11+E19						

Data Modules



Data Modules





Toolkit

- Plug-In for Gnumeric
  - Open source on Linux Platforms
- Prototype implementation
- Supports analysis for
  - Logical areas,
  - Semantic Classes
  - Data Modules
- Builds these abstractions, but no further processing is done

Discussion

- Not always helpful
  - Cannot check numbers for their adequacy
  - Aims for certain kinds of spreadsheets
  - Consistent usage of the wrong formula might not be detected
- Parametrization for semantic classes is a limitation
- Inspection of Data Modules still tedious, but less effort than inspecting a DDG

Discussion

- Scalable, because
  - Abstract representation is far compacter than the original
  - Small changes will lead to significantly different results
  - Spreadsheet with 1200 formula cells – 23 semantic classes
  - DDG with 68 nodes – 5 Data Modules

Questions? Comments?

markus.clermont@ul.ie