Towards a Catalog of Spreadsheet Smells

Jácome Cunha¹, João P.Fernandes¹,², Hugo Ribeiro¹, João Saraiva¹

¹ HASLab / INESC TEC & Universidade do Minho, Portugal
² Universidade do Porto, Portugal

ICCSA 2012
Why do Spreadsheets matter?

- 95% of all U.S. firms
- 90% of all analysts in industry
- 50% of all SSs are the basis for decisions

Omnipresent, Easy-to-use, Multi-purpose, Flexible
Economy losses of $10 billion/year!

Due to a calculating error by their financial advisors, West Baraboo officials learned Thursday they will be paying about $400,000 more over the lifetime of their most recent 10-year borrowing plan than originally projected.

During its regular December meeting, the West Baraboo Village Board looked back over last month's decision to sell $1.1 million in general obligation bonds to cover a variety of village projects, said Village Clerk Mary Klingenmeyer. The review was required after the board received a letter from its financial advisory firm, Ehlers of Brookfield.

Ehlers advisor James Mann said "operator error" resulted in a spreadsheet underestimating the total cost of the 10-year bond.

"We committed to fund [enrollment] growth and this is an important part of growth," said Senate Budget Chairman Lyde Hillyard on Wednesday. "We would hope to get it fixed, and I think that's going to be our first priority."

The $25 million represents less than 1 percent of the state's overall $3 billion-plus education budget.

State Superintendent Larry Shumway attributed the mistake to "a faulty reference" in a spreadsheet. He emphasized that no money was misappropriated. He called the mistake "significant" but "manageable."
Inquérito da PGR atinge jornalistas

Publicado em 2006-02-16
CARLOS VARELA *

Magistrados judiciais e elementos da Polícia Judiciária entraram ontem de rompante na redacção do jornal "24 Horas", em Lisboa, numa operação-surpresa para apreensão de material que o Ministério Público considera associado ao caso do "envelope 9" e ao processo "Casa Pia", que deu origem a uma notícia publicada há mais de um mês naquele diário. A PJ realizou também uma busca, em Portalegre, à casa do jornalista free-lancer co-autor do trabalho.

Na notícia, publicada a 13 de Janeiro, o jornal dava conta da existência de um conjunto de registos telefónicos de altas figuras do Estado, entre as quais Jorge Sampaio e Souto Moura, que surge agregado ao processo Casa Pia, com base num ficheiro Excel enviado pela PT, englobado no famoso "envelope 9".
Many more horror stories...

EuSpriG Original Horror Stories

http://www.eusprig.org/stories.htm
In fact Spreadsheets lack advanced mechanisms found in modern programming languages, namely:

- *Type and Modular Systems*
- *Abstractions*
- *Debuggers/Profilers*
- *Best Practice Rules...*
This paper

- We propose a catalog of bad smells for spreadsheets
- We implement the catalog as the SmellSheet Detective Tool
- We refine the original catalog by executing our tool with the large EUSES spreadsheet repository and by (manually) evaluate the results.
Code Smell

Code Smell, or Bad Smell, as proposed by Martin Fowler, is a concrete evidence that a piece of sw may have a problem.

(usually, a smell is not an error!)
Based on our personal experiences, we propose an initial catalog of spreadsheet bad smells

**Type Smells**
- Empty cell
- Patterns

**Content Smells**
- String distance
- Reference to empty cells

**Quasi-Functional Dependencies**

**Statistical Smells**
- Std Deviation
We validate our initial catalog against real spreadsheets: the EUSES repository (5606 spreadsheets, organized in 6 categories)
SpreadSheet Smells: Catalog Validation

Results obtained from 180 EUSES Spreadsheets

<table>
<thead>
<tr>
<th>Smell/Category</th>
<th>Database</th>
<th>Financial</th>
<th>Grades</th>
<th>Homework</th>
<th>Inventory</th>
<th>Modeling</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empty cells</td>
<td>53</td>
<td>27</td>
<td>56</td>
<td>42</td>
<td>109</td>
<td>106</td>
<td>393</td>
</tr>
<tr>
<td>Patterns</td>
<td>86</td>
<td>62</td>
<td>66</td>
<td>58</td>
<td>111</td>
<td>114</td>
<td>497</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>33</td>
<td>35</td>
<td>134</td>
<td>14</td>
<td>32</td>
<td>7</td>
<td>255</td>
</tr>
<tr>
<td>String Dist.</td>
<td>25</td>
<td>3</td>
<td>1357</td>
<td>231</td>
<td>158</td>
<td>658</td>
<td>2442</td>
</tr>
<tr>
<td>CFD</td>
<td>12</td>
<td>13</td>
<td>150</td>
<td>2</td>
<td>13</td>
<td>24</td>
<td>204</td>
</tr>
<tr>
<td>Ref2empty</td>
<td>0</td>
<td>23</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>209</strong></td>
<td><strong>163</strong></td>
<td><strong>1790</strong></td>
<td><strong>347</strong></td>
<td><strong>423</strong></td>
<td><strong>909</strong></td>
<td><strong>3841</strong></td>
</tr>
</tbody>
</table>
We manually inspect the 180 spreadsheets and empirically classify the computed smells as:

- **Low Smells**
- **Medium Smells**
- **High Smells**
- **Not Smells**
SpreadSheet Smells: Catalog Evaluation

<table>
<thead>
<tr>
<th>Smell</th>
<th>Level</th>
<th>Low smells</th>
<th>Medium smells</th>
<th>High smells</th>
<th>Not smells</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empty cells</td>
<td></td>
<td>115</td>
<td>4</td>
<td>0</td>
<td>274</td>
<td>393</td>
</tr>
<tr>
<td>Patterns</td>
<td></td>
<td>124</td>
<td>9</td>
<td>0</td>
<td>364</td>
<td>497</td>
</tr>
<tr>
<td>Std. Dev. cells</td>
<td></td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>234</td>
<td>255</td>
</tr>
<tr>
<td>String Dist.</td>
<td></td>
<td>13</td>
<td>2</td>
<td>7</td>
<td>2420</td>
<td>2442</td>
</tr>
<tr>
<td>QFD</td>
<td></td>
<td>64</td>
<td>13</td>
<td>6</td>
<td>121</td>
<td>204</td>
</tr>
<tr>
<td>Ref2empty</td>
<td></td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>24</td>
<td>50</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td></td>
<td><strong>346</strong></td>
<td><strong>37</strong></td>
<td><strong>21</strong></td>
<td><strong>3437</strong></td>
<td><strong>3841</strong></td>
</tr>
</tbody>
</table>

90% of the smelly cells detected were not confirmed...
We consider three types of refinement:

1- *Overlapped Smells*
   Empty cells are also detected by patterns...

2- *New Smells Detected*
   Summation cells as constants!!
   Layout smells...

3- *Wrong Smells*
   *String distance alg. needs improvements*
   *numeric cells*
### SpreadSheet Smells: Catalog Refinement

<table>
<thead>
<tr>
<th>Smell</th>
<th>Low smells</th>
<th>Medium smells</th>
<th>High smells</th>
<th>Not smells</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empty cells</td>
<td>115</td>
<td>4</td>
<td>0</td>
<td>274</td>
<td>393</td>
</tr>
<tr>
<td>Patterns</td>
<td>9</td>
<td>5</td>
<td>0</td>
<td>90</td>
<td>101</td>
</tr>
<tr>
<td>Std. Dev. cells</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>234</td>
<td>255</td>
</tr>
<tr>
<td>String Dist.</td>
<td>13</td>
<td>2</td>
<td>7</td>
<td>290</td>
<td>312</td>
</tr>
<tr>
<td>QFD</td>
<td>64</td>
<td>13</td>
<td>6</td>
<td>121</td>
<td>204</td>
</tr>
<tr>
<td>Ref2empty</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>24</td>
<td>50</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>231</strong></td>
<td><strong>33</strong></td>
<td><strong>21</strong></td>
<td><strong>1033</strong></td>
<td><strong>1315</strong></td>
</tr>
</tbody>
</table>

21.7% of correct smells detected
54 out 231 are Medium or High Smells!
We presented a detailed study on spreadsheet smell detection.

We proposed the first catalog of spreadsheet smells.

We developed the SmellSheet Detective: a tool that implements the proposed catalog.
Available at

http://sssaapp.di.uminho.pt