Testing Interoperability of Office Applications

Miloš Šrámek

Society for Open Information Technologies
Slovakia

This work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License.

Society for Open Information Technologies (Slovakia)
Importance of Interoperability

- Interoperability: application and format independent rendering (of office documents)
- Important when
  - Living in a heterogeneous environment
  - Moving from one application to another
- Expectations:
  - Full application and format independent rendering of office documents
Users' Opinions

- **A frequent one:** MS is always right
- **A not that frequent one:** MS (mis)uses formats to protect its market dominance
- Both are (usually) based on random experience with random documents
- To know the real situation we need a solid framework for interoperability testing
Earlier Work

- Presented at Plugfest 2011 in Berlin:
  - ODF document overlays for a detailed visual inspection
  - An *overlap index* for numeric comparison
  - Conversion tools at [http://www.officeshots.org/](http://www.officeshots.org/) used

<table>
<thead>
<tr>
<th>File</th>
<th>#</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>AbiWord_2_8_6_Linux</td>
<td>0</td>
<td>100</td>
<td>81,92</td>
<td>81,92</td>
<td>24,54</td>
<td>24,55</td>
<td>24,54</td>
<td>24,6</td>
<td>24,6</td>
<td>24,6</td>
<td>24,3</td>
</tr>
<tr>
<td>AbiWord_2_9_0_Linux</td>
<td>1</td>
<td>81,92</td>
<td>100</td>
<td>100</td>
<td>24,51</td>
<td>22,89</td>
<td>22,89</td>
<td>22,95</td>
<td>22,95</td>
<td>22,95</td>
<td>22,71</td>
</tr>
<tr>
<td>AbiWord_2_9_1_Linux</td>
<td>2</td>
<td>81,92</td>
<td>100</td>
<td>100</td>
<td>24,51</td>
<td>22,89</td>
<td>22,89</td>
<td>22,95</td>
<td>22,95</td>
<td>22,95</td>
<td>22,71</td>
</tr>
<tr>
<td>KOffice_2_3_3_Linux_B</td>
<td>3</td>
<td>24,54</td>
<td>24,51</td>
<td>24,51</td>
<td>100</td>
<td>34,44</td>
<td>34,53</td>
<td>34,3</td>
<td>34,3</td>
<td>34,3</td>
<td>33,02</td>
</tr>
<tr>
<td>StarOffice_9_2_Linux</td>
<td>8</td>
<td>24,55</td>
<td>22,89</td>
<td>22,89</td>
<td>34,44</td>
<td>100</td>
<td>99,51</td>
<td>79,01</td>
<td>79,01</td>
<td>79,01</td>
<td>56,29</td>
</tr>
<tr>
<td>LibreOffice_3_3_2_Linux</td>
<td>4</td>
<td>24,54</td>
<td>22,89</td>
<td>22,89</td>
<td>34,53</td>
<td>99,51</td>
<td>100</td>
<td>79,32</td>
<td>79,32</td>
<td>79,32</td>
<td>56,33</td>
</tr>
<tr>
<td>LibreOffice_3_3_3_Linux</td>
<td>5</td>
<td>24,6</td>
<td>22,95</td>
<td>22,95</td>
<td>34,3</td>
<td>79,01</td>
<td>79,32</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>60,8</td>
</tr>
<tr>
<td>LibreOffice_3_4_1_Linux</td>
<td>6</td>
<td>24,6</td>
<td>22,95</td>
<td>22,95</td>
<td>34,3</td>
<td>79,01</td>
<td>79,32</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>60,8</td>
</tr>
<tr>
<td>OpenOffice_org_3_4Beta</td>
<td>7</td>
<td>24,6</td>
<td>22,95</td>
<td>22,95</td>
<td>34,3</td>
<td>79,01</td>
<td>79,32</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>60,8</td>
</tr>
<tr>
<td>Microsoft_Office_2007</td>
<td>9</td>
<td>24,3</td>
<td>22,71</td>
<td>22,71</td>
<td>33,02</td>
<td>56,29</td>
<td>56,33</td>
<td>60,8</td>
<td>60,8</td>
<td>60,8</td>
<td>100</td>
</tr>
<tr>
<td>Microsoft_Office_2010</td>
<td>10</td>
<td>24,3</td>
<td>22,71</td>
<td>22,71</td>
<td>33,02</td>
<td>56,29</td>
<td>56,33</td>
<td>60,8</td>
<td>60,8</td>
<td>60,8</td>
<td>100</td>
</tr>
</tbody>
</table>
The 'DoCmp' Framework

- Compare and grade **document pairs**
- Four independent error measures:
  - **Per page** (we assume single-page documents):
    - Page Height Error (PHE): characterizes errors in line spacing:
      - \( PHE = |h(d1) - h(d2)| \)
    - Line Number Difference (LND): missing lines
  - **Per line**:
    - Feature Distance Error (FDE): maximum distance between features of aligned lines
    - Line Position Error (LPE): horizontal shift of dominating line segments
- Segmentation of documents (as images) in individual lines is required
The Feature Distance Error

- Observation: Maximum value of $|\text{df}(L_1) - \text{df}(L_2)|$ is equal to a distance between different local features
- Characterizes difference on a character level

|Subtraction|

Document line L1

|Document line L2|

DF(L1)  DF(L2)
The Comparison Algorithm

- For both compared documents (pdf-s)
  - Convert the pdf-s in images d1 and d2
  - Crop both images and compute Page Height Error
  - Segment both images in lines and interline spaces
  - Get LND by counting lines in both images
  - For each pair of lines of d1 and d2
    - Align the pairs horizontally and vertically
    - Compute the distance field and find the difference maximum
  - Find FDE as the maximum of the distance field difference maxima for all document lines
  - Find LPE as the maximum of the horizontal misalignments for all document lines
The DoCmp framework: Part 1, preparation of test files

● Prepare **source** test documents in the native format of each tested application (manual)
  ● Tested applications: LO3.6, LO40, LO41, AO34, AO40, MSO10, MSO13, CW27, AW29...)
  ● Native formats:
    ● LO/АOO: ODF
    ● MS20XX: OOXML
  ● Example file: **bullets.LO41.odt** (the name matters)

● Convert the source document to all tested formats by all applications (automatic)
  ● Example file: **bullets.LO41.docx**
  ● Tested formats: odt, docx, rtf, doc
The DoCmp framework: Part 2, printing to pdf

- Print the source and converted documents to pdf (automatic)
  
  - Example 1: `bullets.LO41.odt.LO41.pdf`
    - Read: source file created in LO41 in odt and printed by LO41 to pdf
  
  - Example 2: `bullets.LO41.odt.MS13.pdf`
    - Read: source file created in LO41 in odt and printed by MS13 to pdf
  
  - Example 3: `bullets.LO41.docx.MS13.pdf`
    - Read: source file created in LO41 in odt, converted by LO41 to docx and printed by MS13 to pdf
The framework: Part 3, evaluation of the similarity measures

- Evaluate similarity measures in pairs with the same source pdf file:
  - bullets.LO41.odt.LO41.pdf vs. bullets.LO41.docx.MS13.pdf

- Used measures:
  - Page height error, normalized to full page (PHE)
  - Line number difference (LND)
  - Feature distance error (FDE, per line, maximum of all lines taken)
  - Line position error (LPE, per line, maximum of all lines taken)

- Measures values stored in a csv table
- Repeated for all cases, source and target formats and applications
The framework: Part 4, creation of reports

- Automatically generated report (ODT file, odfpy used):
  - Summaries per format, case and application
  - Detailed data on source target application pairs
    - Tables with grading
    - Images of worst case lines according to the feature distance error

- The used grading scale:

<table>
<thead>
<tr>
<th></th>
<th>Grade 0 Ideal</th>
<th>Grade 1 Very good</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5 Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHE [mm]</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>&gt; 20</td>
</tr>
<tr>
<td>LND</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&gt; 0</td>
</tr>
<tr>
<td>FDE [mm]</td>
<td>0</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>&gt; 4</td>
</tr>
<tr>
<td>LPE [mm]</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>&gt; 8</td>
</tr>
</tbody>
</table>
Visual comparison

- Alternative / supplement to the report
- A combined view of the printed converted and source files. For example:
  - `bullets.LO41.odt.LO41.pdf` vs. `bullets.LO41.docx.MS13.pdf`
- Three types of display:
  - Page overlay view, shows predominantly errors in line spacing
  - Aligned row overlay view, line details well visible
  - Side-by-side view, perhaps useful for user studies
The Page Overlay View

Basic numbering, Font Times new roman

1. one
  1. one one
  1. one one
  1. one one

  1. one one
  1. one one
  1. one one

1. one one
  3. uvwxyz abcd efgh ijkl mnop qrst

2. one two
  2. one two one

3. one three
  3. one three
  1. one three one

2. two one two
The Aligned Row Overlay View

Basic numbering, font times new roman

1. one
   1. one
   1.1. one one one - longer text for second line indent testing ... abcd efgh ijk mnop mnop q rst uvwxyz abcd efgh ijk mnop q rst uvwxyz
   1.2. 2. one one two
   1.3. 3. one one three
   2. one two
   2.1. 1. one two one
   3. one three
   3.1. 1. one three one
   2. two
   1. two one
   1.1. 1. two one one
The Side-by-Side View

**Original**

Basic numbering, Font Times new roman

1. one
   1. one
      1. one one - longer text for second line indent testing ... abcd efgh ijkl mnop qrst uvwxyz abcd efgh ijkl mnop qrst uvwxyz
   2. one one two
   3. one one three
2. one two
   1. one two one
   3. one three
3. one three
   1. one three one
2. two
   1. two one
      1. two one one
   2. two two
      1. two two one
      2. two two two
   3. two three
3. two three

Basic numbering, Font Times new roman

**Test 8**

Basic numbering, Font Times new roman

1. one
   1. one one
   1.1. one one one - longer text for second line indent testing ... abcd efgh ijkl mnop qrst uvwxyz abcd efgh ijkl mnop qrst uvwxyz
   1.2. one one two
   1.3. one one three
2. one two
   2.1. one two one
   3. one three
   3.1. one three one
2. two
   1. two one
   1.1. two one one
   2. two two
   2.1. two two one
   2.2. two two two
   2.3. two two three
3. two three

Basic numbering, Font Times new roman
Implementation

- A bash script for format conversion and printing to pdf (on Windows using Cygwin)
  - Only applications with command line interface:
    - AOO, LO, MSO, GoogleDocs, Abiword, Calligra Words, ...
- Support programs enabling the script-based conversion:
  - MSO (the OfficeConvert tool, C#)
  - AOO (communication with AOO running as a server)
  - GoogleDocs (using Google API)
The Testing Workflow

1. Prepare source documents, store them in a predefined directory structure (all platforms)
2. Run the conversion script (all platforms)
   - Bash, to be repeated on all platforms
3. Run the printing script (all platforms)
   - Bash, to be repeated on all platforms
4. Run the script for evaluation of the measures and preparation of views (bash, python, only Linux)
5. Run the script for report generation (python, only Linux)
6. Enjoy the report and see how good/bad the interoperability is
The Experiment: Test Cases

• One page documents featuring
  • Formatted text
  • Bullets
  • Bullets advanced
  • Numbering
  • Numbering advanced
  • Chapters
  • Chapters numbered
  • Tables
  • Tables advanced
  • Image insertion
# Numeric Results for the 'Numbering' Test Case

<table>
<thead>
<tr>
<th>ODT</th>
<th>AO40</th>
<th>LO40</th>
<th>LO41</th>
<th>MS13</th>
</tr>
</thead>
<tbody>
<tr>
<td>AO40</td>
<td>[0, 0, 0, 0]</td>
<td>[0, 0, 0, 0]</td>
<td>[0, 0, 0, 0]</td>
<td>[0, 1, 1, 0]</td>
</tr>
<tr>
<td>LO40</td>
<td>[0, 0, 0, 0]</td>
<td>[0, 0, 0, 0]</td>
<td>[0, 1, 0, 0]</td>
<td>[0, 1, 1, 0]</td>
</tr>
<tr>
<td>LO41</td>
<td>[0, 1, 0, 0]</td>
<td>[0, 1, 0, 0]</td>
<td>[0, 0, 0, 0]</td>
<td>[0, 1, 1, 0]</td>
</tr>
<tr>
<td>MS13</td>
<td>[0, 1, 1, 0]</td>
<td>[0, 1, 1, 0]</td>
<td>[0, 1, 1, 0]</td>
<td>[0, 0, 0, 0]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DOC</th>
<th>AO40</th>
<th>LO40</th>
<th>LO41</th>
<th>MS13</th>
</tr>
</thead>
<tbody>
<tr>
<td>AO40</td>
<td>[0, 0, 0, 0]</td>
<td>[0, 0, 0, 0]</td>
<td>[0, 0, 0, 0]</td>
<td>[3, 1, 1, 5]</td>
</tr>
<tr>
<td>LO40</td>
<td>[0, 0, 0, 0]</td>
<td>[0, 0, 0, 0]</td>
<td>[0, 1, 0, 0]</td>
<td>[2, 1, 1, 5]</td>
</tr>
<tr>
<td>LO41</td>
<td>[0, 1, 0, 0]</td>
<td>[0, 1, 0, 0]</td>
<td>[0, 0, 0, 0]</td>
<td>[2, 1, 1, 5]</td>
</tr>
<tr>
<td>MS13</td>
<td>[0, 1, 1, 0]</td>
<td>[0, 1, 1, 0]</td>
<td>[0, 1, 1, 0]</td>
<td>[0, 0, 0, 0]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DOCX</th>
<th>AO40</th>
<th>LO40</th>
<th>LO41</th>
<th>MS13</th>
</tr>
</thead>
<tbody>
<tr>
<td>AO40</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LO40</td>
<td>[0, 5, 5, 0]</td>
<td>[0, 0, 0, 0]</td>
<td>[0, 1, 0, 0]</td>
<td>[5, 1, 1, 0]</td>
</tr>
<tr>
<td>LO41</td>
<td>[0, 5, 5, 0]</td>
<td>[0, 1, 0, 0]</td>
<td>[0, 0, 0, 0]</td>
<td>[5, 1, 1, 0]</td>
</tr>
<tr>
<td>MS13</td>
<td>[0, 5, 5, 0]</td>
<td>[0, 1, 1, 0]</td>
<td>[0, 1, 1, 0]</td>
<td>[0, 0, 0, 0]</td>
</tr>
</tbody>
</table>

- Poor grades should be analyzed visually
'Numbering': Visual evaluation

- ODT, LO40→LO41: grade 1
- DOCX, LO41→AO40: grade 5
  
  1. one one one - longer text for second line indent testing

- DOCX, MS13→AO40: grade 5
  
  1.1. 1. one one one - longer text for second line indent testing

- DOCX, LO40, LO41→MS13: grade 5

- DOC, AO40, LO40, LO41→MS13: grade 5
  
  - The last line missing in MS13
## Summaries

- **ALL cases, LO41, AO40 and MS13:**

<table>
<thead>
<tr>
<th>format</th>
<th>text height err.</th>
<th>feature dist. err.</th>
<th>line position err.</th>
<th>line number diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc</td>
<td>0.9</td>
<td>1.5</td>
<td>1.3</td>
<td>1.1</td>
</tr>
<tr>
<td>docx</td>
<td>1.1</td>
<td>2.0</td>
<td>1.9</td>
<td>0.0</td>
</tr>
<tr>
<td>odt</td>
<td>1.0</td>
<td>1.6</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- **Worst grades:**

<table>
<thead>
<tr>
<th>format</th>
<th>text height err.</th>
<th>feature dist. err.</th>
<th>line position err.</th>
<th>line number diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>docx</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>0.0</td>
</tr>
<tr>
<td>odt</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- **All cases, ODT**

<table>
<thead>
<tr>
<th>source/target</th>
<th>LO41</th>
<th>AO40</th>
<th>MS13</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO41</td>
<td>0.0</td>
<td>0.4</td>
<td>1.8</td>
</tr>
<tr>
<td>AO40</td>
<td>0.5</td>
<td>0.0</td>
<td>1.9</td>
</tr>
<tr>
<td>MS13</td>
<td>1.5</td>
<td>1.4</td>
<td>0.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>source/target</th>
<th>LO41</th>
<th>AO40</th>
<th>MS13</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO41</td>
<td>0.0</td>
<td>1.0</td>
<td>5.0</td>
</tr>
<tr>
<td>AO40</td>
<td>5.0</td>
<td>0.0</td>
<td>5.0</td>
</tr>
<tr>
<td>MS13</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
</tr>
</tbody>
</table>
Summary: LO41 & AO40 vs. MS13

- **LO41 vs. MS13:**
  
  **Average grades:**
  
<table>
<thead>
<tr>
<th>format</th>
<th>text height err.</th>
<th>feature dist. err.</th>
<th>line position err.</th>
<th>line number diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc</td>
<td>0.8</td>
<td>1.5</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>docx</td>
<td>1.3</td>
<td>1.2</td>
<td>1.1</td>
<td>0.0</td>
</tr>
<tr>
<td>odt</td>
<td>1.0</td>
<td>1.7</td>
<td>1.2</td>
<td>0.0</td>
</tr>
</tbody>
</table>

  **Worst grades:**
  
<table>
<thead>
<tr>
<th>format</th>
<th>text height err.</th>
<th>feature dist. err.</th>
<th>line position err.</th>
<th>line number diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc</td>
<td>3.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>docx</td>
<td>5.0</td>
<td>3.0</td>
<td>5.0</td>
<td>0.0</td>
</tr>
<tr>
<td>odt</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- **AO40 vs. MS13:**
  
  **Average grades:**
  
<table>
<thead>
<tr>
<th>format</th>
<th>text height err.</th>
<th>feature dist. err.</th>
<th>line position err.</th>
<th>line number diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc</td>
<td>0.9</td>
<td>1.4</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>docx</td>
<td>0.6</td>
<td>1.7</td>
<td>1.4</td>
<td>0.0</td>
</tr>
<tr>
<td>odt</td>
<td>1.2</td>
<td>1.6</td>
<td>1.2</td>
<td>0.0</td>
</tr>
</tbody>
</table>

  **Worst grades:**
  
<table>
<thead>
<tr>
<th>format</th>
<th>text height err.</th>
<th>feature dist. err.</th>
<th>line position err.</th>
<th>line number diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>docx</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>0.0</td>
</tr>
<tr>
<td>odt</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

(tested both reading and writing)
Shortcomings and Future Work

- The error measures are perhaps not optimal
- The grading scale is perhaps not optimal
- Correlation between the proposed approach and user perception is unknown
  - Run a user study?
- Can the results be used for LO marketing?
  - Results published at independent site, e.g. Phoronix?
- Where to publish the code?
  - Github?
Download

- The code and test cases: https://gitorious.org/docmp
- The test cases presented in Milano, converted and printed files (52MB): http://ubuntuone.com/0sBUtiZdFyTYeI6VM1Wcal
- The test cases presented in Milano, reports: http://ubuntuone.com/2nlBtaXkY8flQtvFOMrbW5

Thank You!
Is an implementation correct?

- Tests show differences, but do not tell us what is correct.
- Assumption:
  - An application may fail in many ways, but correct is only in one way.
- Conclusion:
  - If two independent applications display one file in the same way, then
    - The file is correct.
    - Both applications can read the file correctly.
    - The application which created the file wrote it correctly.