From RTF to XML to LaTeX

Andre Dierker Arne Jans
Stephan Lehmke
QuinScape GmbH, Thomasstraße 1, 44135 Dortmund, Germany
{Andre.Dierker,Arne.Jans,Stephan.Lehmke}@QuinScape.de
http://www.QuinScape.de

Februar 26, 2005

Abstract

This paper shows how the widely used Rich Text Format (originally specified by Microsoft) can be processed to produce XML and how the resulting XML-file can be processed by LaTeX to produce print-quality PDF-files. To convert the RTF-files to a specific XML-format we use the open-source-tool Majix which can be found on Sourceforge.

We then take XMLTeX to process the generated XML-file. We had to create an output that is as near as possible to Word's. An effort was made to reach this goal even with constructs such as lists, tabstops and especially tables.

Using XML as an intermediary format in typesetting RTF has the advantage that structural transformations are much easier based on XML even if the XML 'only' reproduces the RTF as faithfully as possible.

Filtering or transforming certain objects or attributes and even correcting typesetting errors can be done by appropriate transformations of the XML files.

1 Introduction

For a commercial project it was necessary to typeset larger documents from automatically generated XML data with embedded references to external RTF files. It was decided to first transform the RTF to XML and include that into the existing XML-structure, all together then being typeset by XMLTeX.

For this, the open-source-tool Majix was extended to achieve the RTF-XML translation. For typesetting the resulting XML an appropriate implementation using XMLTeX was created which will be uploaded to CTAN eventually, providing another open-source way of handling RTF with TeX.

2 Results

We begin with a RTF-File. As you can see in figure 2 OpenOffice isn't able to handle recorded changes to the file correctly. For example in 'Betrag für 20042 Tsd EUR' (first row, second column) the '20042' is meant to be a year. Originally is was 2002 but the last '2' was deleted and replaced with '4'. However the '2' remained as old version in the RTF and and is tagged as 'deleted'. In opposite to Majix OpenOffice doesn't recognise the responsible control word. The corresponding RTF-Code is shown in the follwing listing. We have selected a quite readable portion. When it comes to font management RTF isn't readable at all.

The RTF-Code looks like this

```
\s16\qj\sa40\widctlpar\adjustright \f18\fs16\lang1031\cgrid {\expnd0\expndtw-2 Anteil, der
aufgrund Artikel 9 des Verwaltungsabkommens vom 5.\~9.\~1957 i. d. F. vom 28.\~2.\~1991 zwi
schen Bund und L\'e4ndern \'fc
ber die Err
ichtung eines Wissenschaftsrats im Haushaltsjahr 1994 voraussichtlich entf\'e4llt.
\par }\pard
\plain \s18\qc\sa40\widctlpar\adjustright \f18\fs16\lang1031\cgrid {\'dcbersicht \'fcber di
e Einnahmen und Ausgaben\line des Wissenschaftsrates
\par }\trow
d \clvertalt\clbrdrt\brdrs\brdrw20 \clbrdrb\brdrs\brdrw20 \cltxlrtb \cellx1021\clvertalt\cl
brdrt\brdrs\brdrw20 \clbrdrb\brdrs\brdrw20 \clbrdrr\brdrs\brdrw20 \cltxlrtb \cellx2042\clve
```

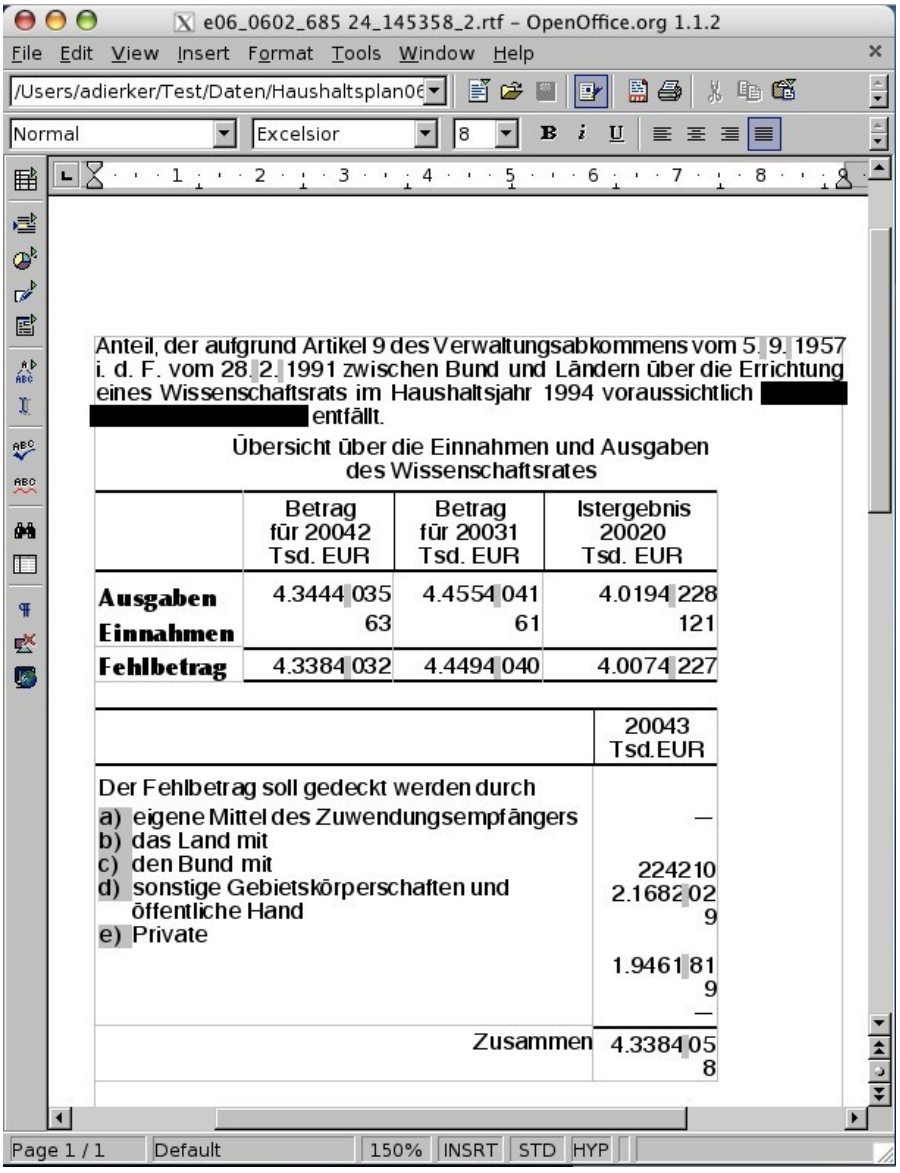


Figure 1: We first have a RTF file. Note the mixed up old and new version of the years and amounts. (Please ignore the font type. The used one wasn't installed and had to be substituted by OpenOffice)

```
rtalt\clbrdrt\brdrs\brdrw20 \clbrdrb\brdrs\brdrw20 \cltxlrtb
\cellx3063\clvertalt\clbrdrt\brdrs\brdrw20 \clbrdrl\brdrs\brdrw20 \clbrdrb\brdrs\brdrw20 \c
ltxlrtb \cellx4253\pard\plain \qc\sb40\sa40\widctlpar\intbl\adjustright \f18\fs16\lang1031\
cgrid {\cell Betrag\line f\'fcr 200}{\revised\revauth1\revdttm1182249745
4}{\deleted
\revauthdel1\revdttmdel1182249745 2}{\line Tsd. EUR\cell }\pard \qc\fi1\li-1\sb40\widctlpar
\intbl\adjustright {Betrag\line f\'fcr 200}{\revised\revauth1\revdttm1182249745 3}{\deleted
\revauthdel1\revdttmdel1182249745 1}{\line Tsd. EUR\cell
Istergebnis
\line 200}{\revised\revauth1\revdttm1182249745 2}{\deleted\revauthdel1\revdttmdel1182249745
 0}{\line Tsd. EUR\cell }\pard \widctlpar\intbl\adjustright {\row }\trowd \clvertalt\cltxlr
tb \cellx1021\clvertalt\cltxlrtb \cellx2042
\clvertalt\
cltxlrtb \cellx3063\clvertalt\cltxlrtb \cellx4253\pard\plain \s1\sb80\sa40\keepn\widctlpar\
intbl\outlinelevel0\adjustright \b\fs20\lang1031\cgrid {\f18\fs16 Ausgaben}
```

After the conversion by Majix we get the XML-Code shown in the next listing. Majix did a great job in giving the data a meaningful structure.

The generated XML is much more readable

```
<par align="justified">
  <tabdeflist>
    <tabdef type="default" align="left" position="12.49mm"/>
  </tabdeflist>
  <parcontent>
    Anteil, der aufgrund Artikel 9 des Verwaltungsabkommens vom 5. 9. 1957 i. d. F.
    vom 28. 2. 1991 zwischen Bund und Ländern über die Errichtung eines
     Wissenschaftsrats im Haushaltsjahr 1994 voraussichtlich entfällt.
  </parcontent>
</par>
<par align="center">
  <tabdeflist>
    <tabdef type="default" align="left" position="12.49mm"/>
  </tabdeflist>
  <parcontent>
    Übersicht über die Einnahmen und Ausgaben<linebreak/>des Wissenschaftsrates
  </parcontent>
</par>
<table>
  <tbody>
    <tr>
      <td width="18.0093mm" valign="top" border-top="0.0pt"
          border-bottom="1.0pt" border-left="0.0pt">
      </td>
      <td width="18.0093mm" valign="top" border-top="0.0pt"
          border-bottom="1.0pt" border-right="1.0pt">
        <par align="center">
          <tabdeflist>
            <tabdef type="default" align="left" position="12.49mm"/>
          </tabdeflist>
          <parcontent>
            Betrag<linebreak/>für 2004<linebreak/>Tsd. EUR
          </parcontent>
        </par>
      </td>
      <td width="18.0093mm" valign="top" border-top="0.0pt"
          border-bottom="1.0pt">
        <par align="center">
          <tabdeflist>
            <tabdef type="default" align="left" position="12.49mm"/>
          </tabdeflist>
          <parcontent>
            Betrag<linebreak/>für 2003<linebreak/>Tsd. EUR
          </parcontent>
        </par>
      </td>
      ...
    </tr>
```

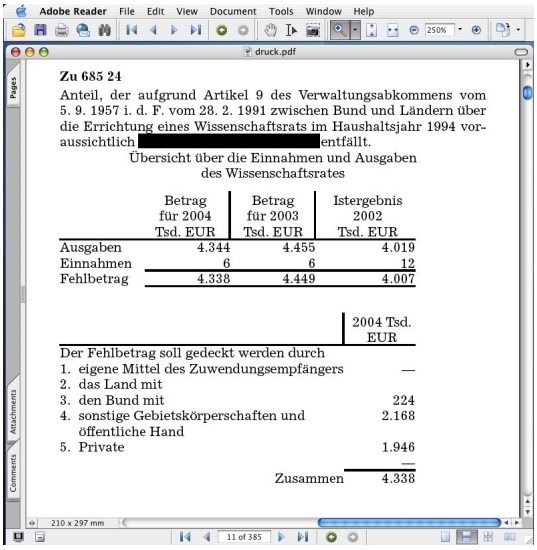
The resulting PDF is shown in figure 2. As you can see there are some differences between the OpenOffice and the LaTeX version. These are the result of some filtering on the XML data demanded by our client.

3 Filtering

Having a well-formed XML one can easily filter the data. For example harmonizing the indentation of unordered lists can be done by deleting the necessary attributes in the XML and providing corresponding defaults.

Another class of transformations deals with the enrichment of semantics by converting visual markup to logical markup: one could search for special XML-constructs perhaps with specific attributes and/or contents and replace them with a 'meaningful' structure.

A good example for this are the so called VEGrids. These are special tables which always have the same layout. A typical VEGrid is shown in figure 3. The headline is always the same as is the tablehead. In the first column there are years, the bottom row is a sum. Usually VEGrids are already given in a XML-structure but sometimes it seems the person dealing with the case doesn't use the right program to input the data but uses Word to create a RTF file with a VE-lookalike. After the XML conversion we get a noname-table as shown in



Zu 685 24

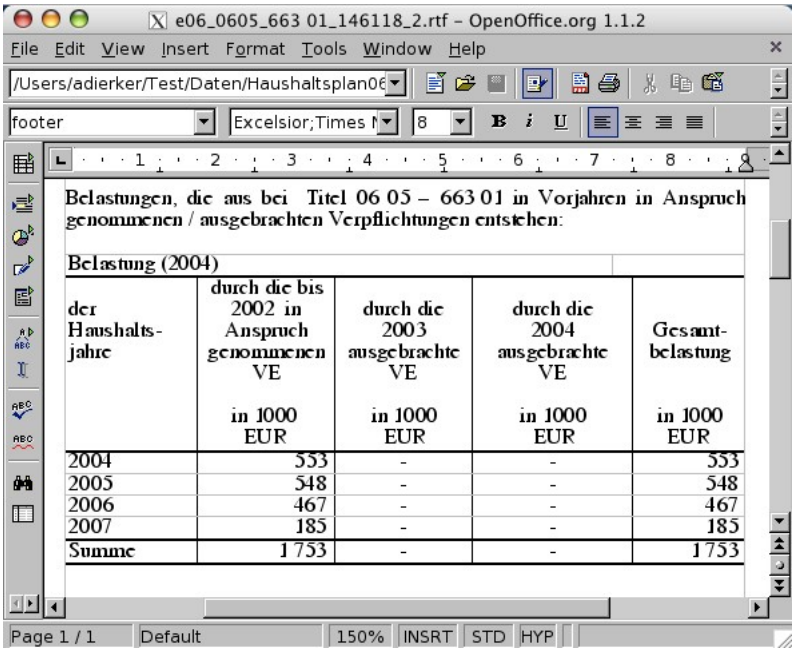
Anteil, der aufgrund Artikel 9 des Verwaltungsabkommens vom 5. 9. 1957 i. d. F. vom 28. 2. 1991 zwischen Bund und Ländern über die Errichtung eines Wissenschaftsrats im Haushaltsjahr 1994 voraussichtlich [redacted] entfällt.

Übersicht über die Einnahmen und Ausgaben des Wissenschaftsrates

	Betrag für 2004 Tsd. EUR	Betrag für 2003 Tsd. EUR	Istergebnis 2002 Tsd. EUR
Ausgaben	4.344	4.455	4.019
Einnahmen	6	6	12
Fehlbetrag	4.338	4.449	4.007

	2004 Tsd. EUR
Der Fehlbetrag soll gedeckt werden durch	
1. eigene Mittel des Zuwendungsempfängers	—
2. das Land mit	
3. den Bund mit	224
4. sonstige Gebietskörperschaften und öffentliche Hand	2.168
5. Private	1.946
Zusammen	4.338

Figure 2: After processing with XMLTEX



Belastungen, die aus bei Titel 06 05 – 663 01 in Vorjahren in Anspruch genommenen / ausgebrachten Verpflichtungen entstehen:

Belastung (2004)

der Haushaltsjahre	durch die bis 2002 in Anspruch genommenen VE in 1000 EUR	durch die 2003 ausgebrachte VE in 1000 EUR	durch die 2004 ausgebrachte VE in 1000 EUR	Gesamtbelastung in 1000 EUR
2004	553	-	-	553
2005	548	-	-	548
2006	467	-	-	467
2007	185	-	-	185
Summe	1753	-	-	1753

Figure 3: A special table called VEGrid in a RTF file

the next listing. Because of the constant layout we could identify these lookalikes and exchange them by 'real' VEGrids. This is shown in the last listing on the next page.

The generated XML (note the numerous omissions)

```
<table>
  <tbody>
    <tr border-bottom="0.5pt" border-left="0.5pt" border-right="0.5pt"
        border-horizontal="0.5pt" border-vertical="0.5pt">
      <td width="72.5135mm" valign="top" border-top="0.0pt"
          border-left="0.0pt" colspan="4">
        <par align="left">
          <tabdeflist>
            <tabdef type="default" align="left" position="12.49mm"/>
          </tabdeflist>
          <parcontent>Belastung (2004) </parcontent></par></td>
```

```
      <td width="17.4978mm" valign="top" border-top="0.0pt"
          border-right="0.0pt" colspan="2"></td>
    </tr>
    <tr border-top="0.5pt" border-bottom="0.5pt" border-left="0.5pt" ...>
      <td width="17.5154mm" valign="top" border-top="1.0pt" ...>
          ...
          <parcontent>
            <linebreak/>der<linebreak/>Haushalts-<linebreak/>jahre
          </parcontent></par></td>
      ...
    </tr>
    <tr border-top="0.5pt" border-bottom="0.5pt" border-left="0.5pt"...>
      <td width="17.5154mm" valign="top" border-top="1.0pt"...>
          ...
          <parcontent>2004</parcontent></par></td>
      <td width="18.1151mm" valign="top" border-top="1.0pt"...>
          ...
          <parcontent>553 </parcontent></par></td>
      <td width="18.1328mm" valign="top" border-top="1.0pt"...>
          ...
          <parcontent>-</parcontent></par></td>
    ...
    </tr>
    <tr border-top="0.5pt" border-bottom="0.5pt" border-left="0.5pt"...>
      <td width="17.5154mm" valign="top" border-top="1.0pt"...>
          ...
          <parcontent>Summe</parcontent></par></td>
      <td width="18.1151mm" valign="top" border-top="1.0pt"...>
          ...
          <parcontent>1 753 </parcontent></par></td>
      ...
    </tr>
  </tbody>
</table>
```

The enriched and filtered XML

```
        <VEGRID>
          <VEROW>
            <VECOLUMN1>2005</VECOLUMN1>
            <VECOLUMN2>553</VECOLUMN2>
            <VECOLUMN3>null</VECOLUMN3>
            <VECOLUMN4>null</VECOLUMN4>
            <VECOLUMN5>553</VECOLUMN5>
          </VEROW>
          <VEROW>
            <VECOLUMN1>2004</VECOLUMN1>
            <VECOLUMN2>548</VECOLUMN2>
            <VECOLUMN3>null</VECOLUMN3>
            <VECOLUMN4>null</VECOLUMN4>
            <VECOLUMN5>548</VECOLUMN5>
          </VEROW>
          ...
          <VESUMROW>
            <VECOLUMN1>Summe</VECOLUMN1>
            <VECOLUMN2>1753</VECOLUMN2>
            <VECOLUMN3>null</VECOLUMN3>
            <VECOLUMN4>null</VECOLUMN4>
            <VECOLUMN5>1753</VECOLUMN5>
          </VESUMROW>
        </VEGRID>
```

4 Further Development

Because there was no need up to now we ignore changes of font size and type. Besides that the management of colours wasn't implemented yet. Perhaps these features will be implemented in future.

As said before it is planned to release the code of the package on CTAN. The extended version of the RTF converter Majix is already available via SourceForge.