Save as XDiML II

The Package for Electronic Theses

X Dissertation Markup Language

Sabine Henneberger, Matthias Schulz, Humboldt University 24.09.2004

Agenda

- Speakers
- Electronic Publishing Group of Humboldt University
  - *Projects*
  - *Document Server*
- Electronic Theses with OpenOffice
- From OpenOffice to XdiML
  - *OOo Con2003*
  - *News*
  - *Filter*
  - *Bat*
- Organisation
- Summary
- Q & A

Speakers

Sabine Henneberger
and Matthias Schulz

- are staff members of the Electronic Publishing Group of the Computer and Mediaservice at Humboldt University (Berlin),
- Sabine Henneberger since 2002 and Matthias Schulz since 1997.
- She developed the conversion tools for XDiML.
- Matthias Schulz was involved in the development of XDiML DTD
- Both developed the template dissertation-hu.stw.

shenneberger@cms.hu-berlin.de matthias.schulz.1@cms.hu-berlin.de
http://edoc.hu-berlin.de

Projects

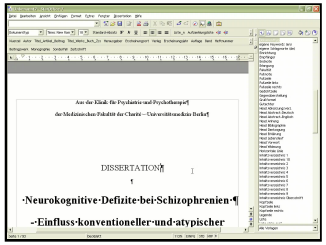
http://edoc.hu-berlin.de/e_projekte_en/

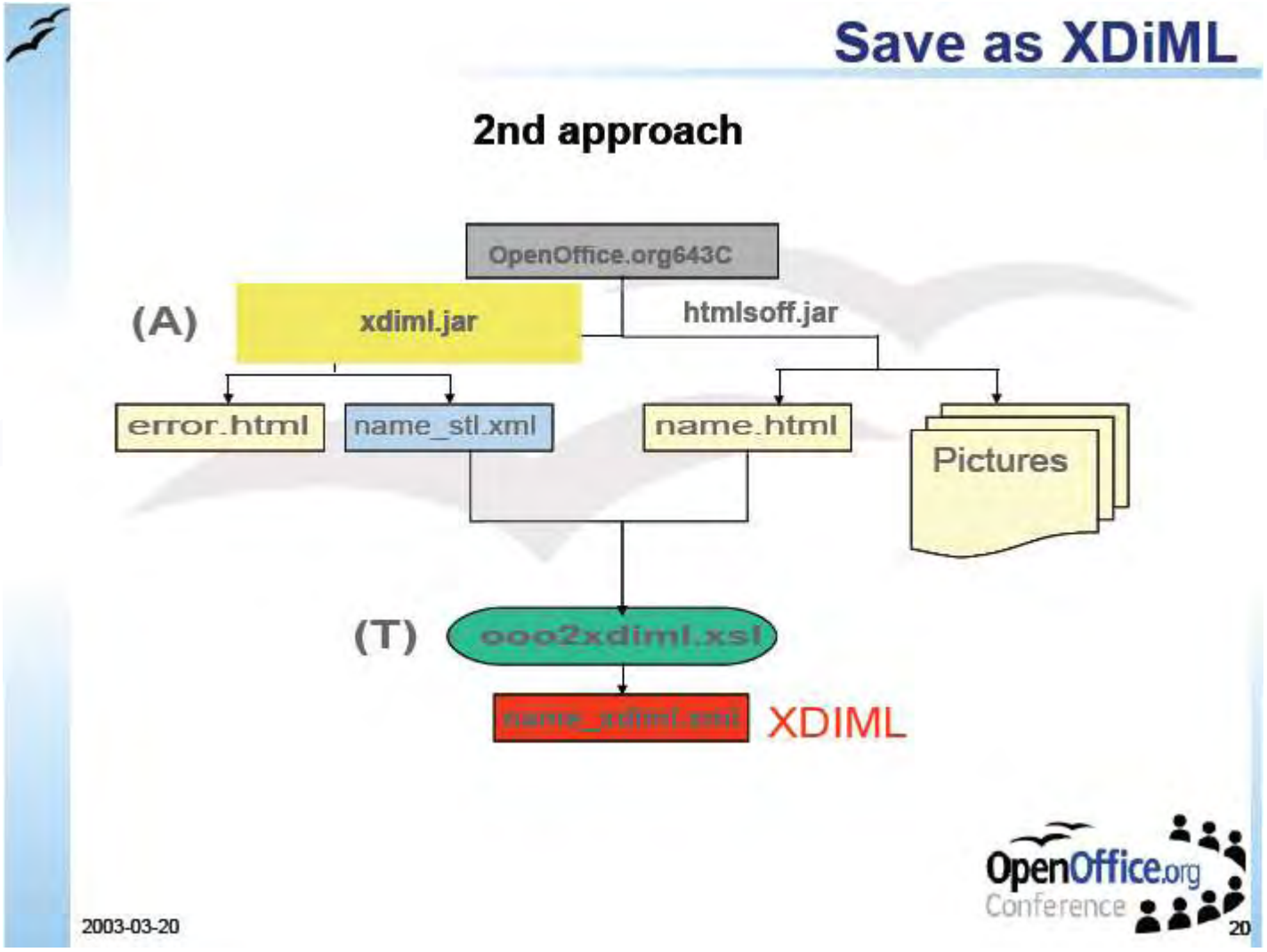
Document Server

http://edoc.hu-berlin.de/

Write your Theses

OOO-Theses-Video




Save as XDiML
2nd approach
OpenOffice.org643C
(A)
xdiml.jar
htmlsoff.jar
error.html
name_stl.xml
name.html
Pictures
(T)
ooo2xdiml.xsl
name_xdiml.xml
XDIML
2003-03-20
OpenOffice.org
Conference
20

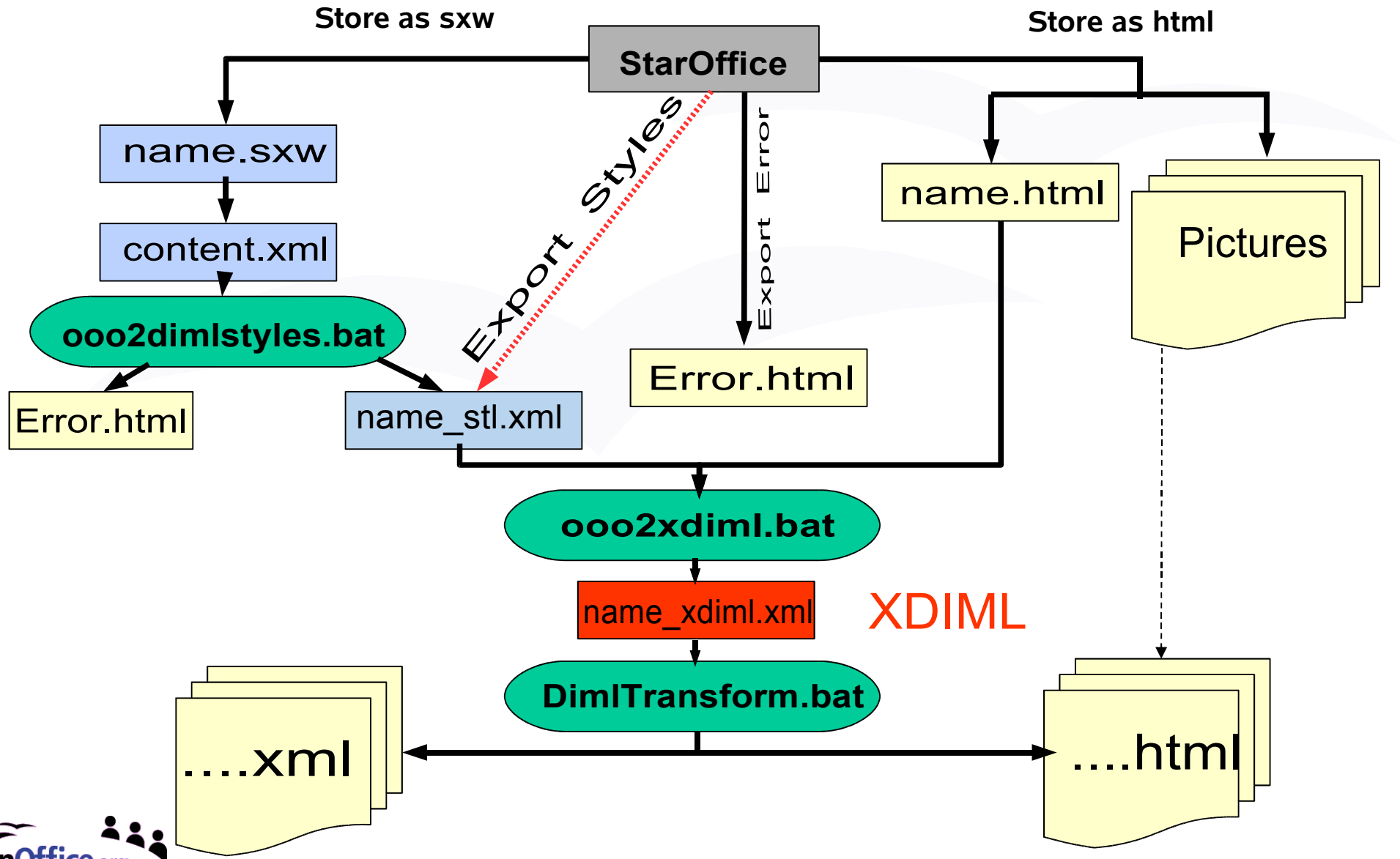
Conclusion

The conversion is a very complex process

- with different tools and several steps
- with different file formats

and no one way process.

Save as XDiML II



content.xml

Structure of content.xml

```
<office:body>
<text:p text:style-name="P1">Analyses of Dairy Cattle Breeding Practices<text:line-break/>in Selected Areas of Ethiopia</text:p>
<text:p text:style-name="P2"/>
<text:p text:style-name="P2"/>
<text:p text:style-name="Dokumenttyp">Dissertation</text:p>
<text:p text:style-name="P3"/>
<text:p text:style-name="Erlangung">zur Erlangung des akademischen Grades doctor rerum agriculturarum<text:line-break/>(Dr. rer. agr.)</text:p>
<text:p text:style-name="P4"/>
<text:p text:style-name="Fakultät">eingerichtet an der Landwirtschaft-Gärtenerischen Fakultät<text:line-break/>der Humboldt-Universität zu Berlin</text:p>
<text:p text:style-name="P4"/>
```

Filters

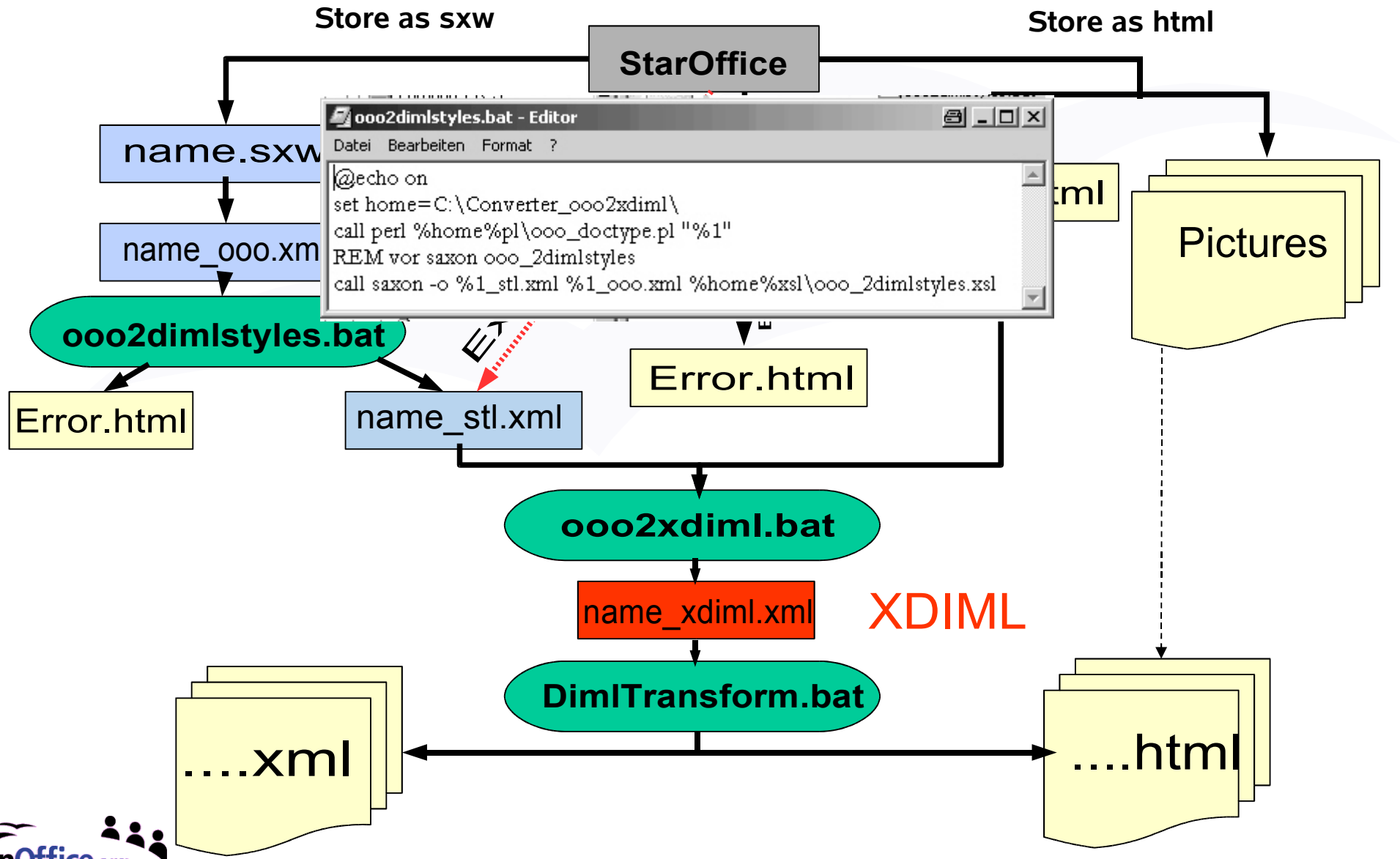
ooo2xdimlstylesfilter.xsl

- delete all elements, we dont need further
- simplify some elements
- convert some characters (symbol) into UTF-8
- change unknown formats to p
- assign our XdiML styles to the text

ooo2error.xsl

- lists all used unknown styles

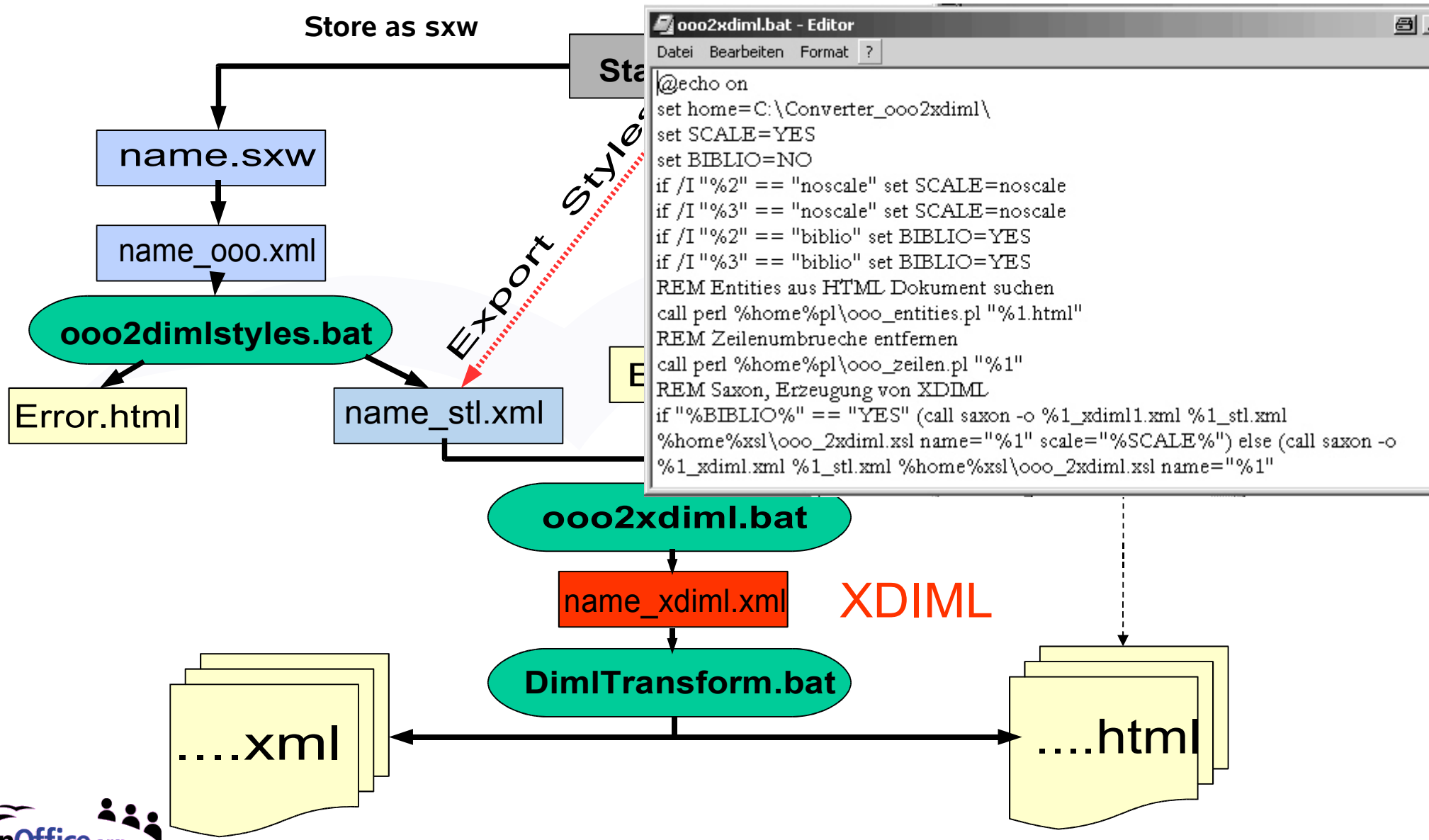
Save as XDiML II



ooo2dimlstyles.bat

```
@echo on
set home=C:\Converter_ooo2xdiml\
call perl %home%pl\ooo_doctype.pl "%1"
REM vor saxon ooo_2dimlstyles
call saxon -o %1_stl.xml %1_ooo.xml
%home%xsl\ooo_2dimlstyles.xsl
```

Save as XDiML II



Save as XDiML II

ooo2xdiml.bat

```
@echo on
set home=C:\Converter_ooo2xdiml\
set SCALE=YES
set BIBLIO=NO
if /I "%2" == "noscale" set SCALE=noscale
if /I "%3" == "noscale" set SCALE=noscale
if /I "%2" == "biblio" set BIBLIO=YES
if /I "%3" == "biblio" set BIBLIO=YES
REM Entities aus HTML Dokument suchen
call perl %home%pl\ooo_entities.pl "%1.html"
REM Zeilenumbrueche entfernen
call perl %home%pl\ooo_zeilen.pl "%1"
REM Saxon, Erzeugung von XDIML
if "%BIBLIO%" == "YES" (call saxon -o %1_xdiml1.xml %1_stl.xml
%home%xsl\ooo_2xdiml.xsl name="%1" scale="%SCALE%") else (call saxon -o
%1_xdiml.xml %1_stl.xml %home%xsl\ooo_2xdiml.xsl name="%1" scale="%SCALE%")
if "%BIBLIO%" == "YES" call perl %home%pl\ooo_bibliography.pl "%1"
```

Save as XDiML II

XML

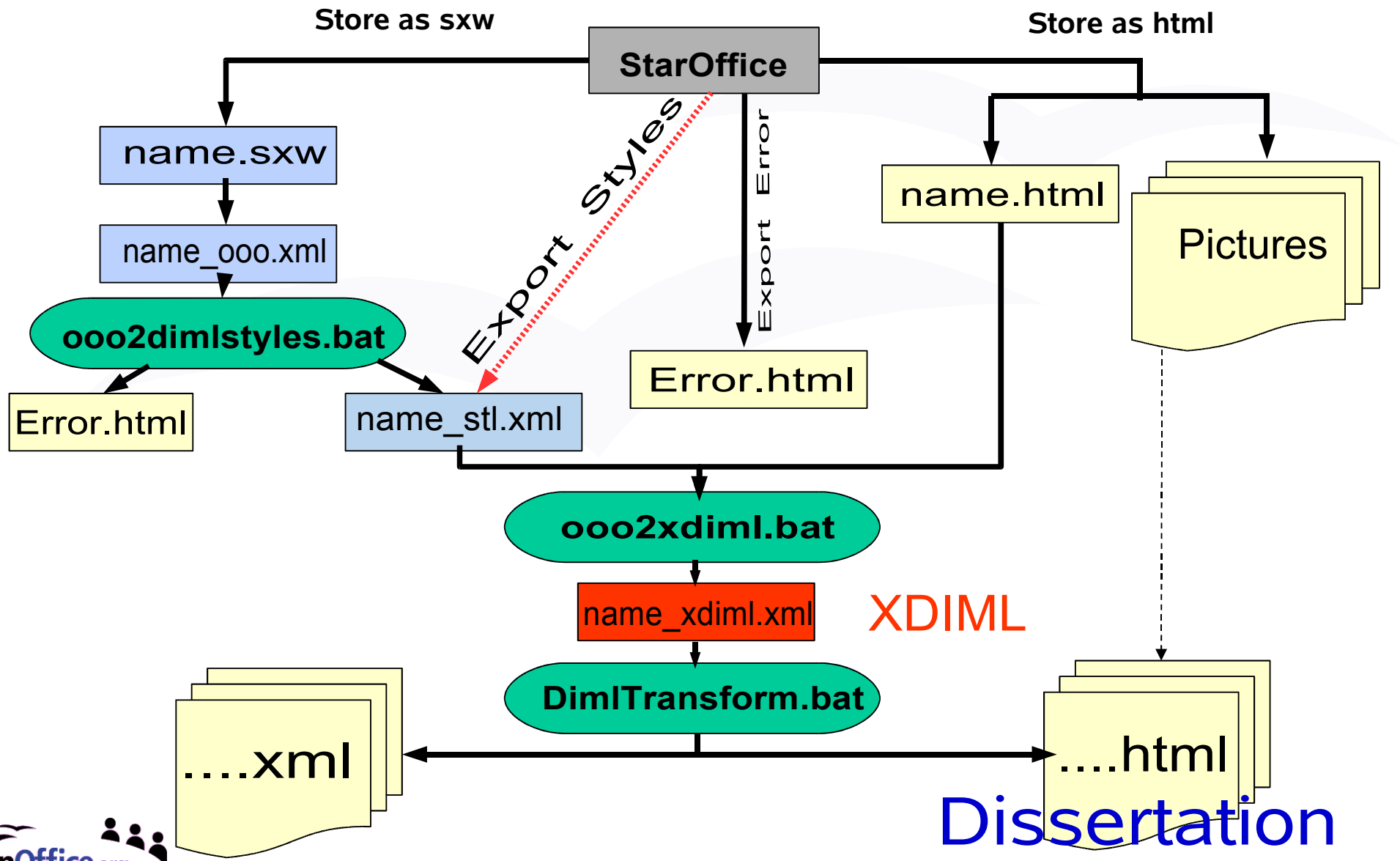
<office:body>

```
<text:h text:style-name="P20" text:level="1">Heading 1</text:h>

<text:h text:style-name="P25" text:level="2">Heading 2 </text:h>

<text:h text:style-name="P26" text:level="3">Heading 3</text:h>

<text:p text:style-name="P27">In the western world, ...</text:p>
```

<etd>

```
<chapter>
  <head>Heading 1</head>
  <section>
    <head> Heading 2</head>
    <subsection>
      <head>Heading 3</head>
      <p>In the western world,...</p>
    </subsection>
  </section>
</chapter>
```

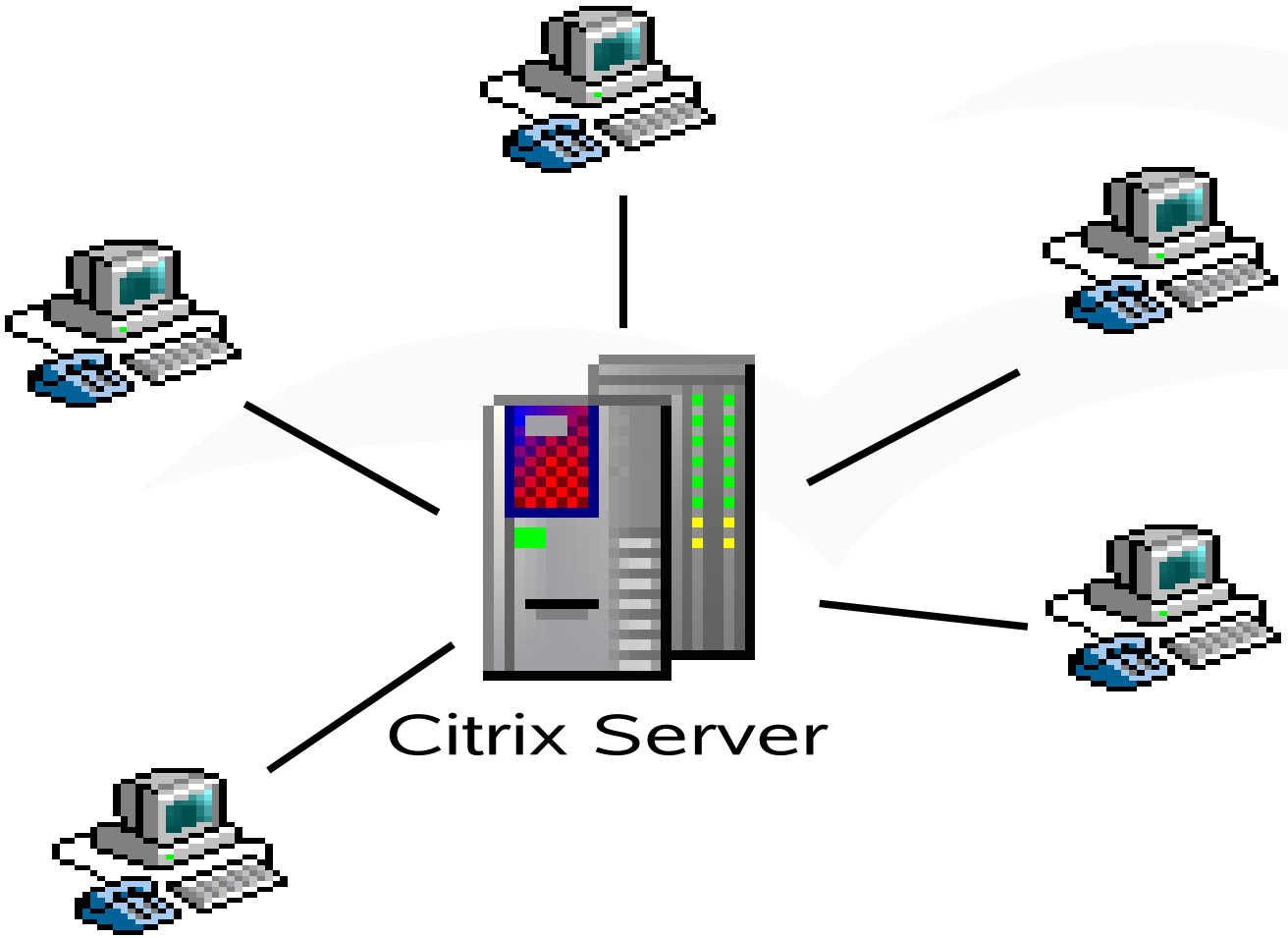
Save as XDiML II



Save as XDiML II

```
...
REM Teilen des stl files in einzelne Bestandteile
saxon -o temp\front.xml %1_stl.xml %home%xsl\ooo_2parts.xsl
REM Konvertieren der einzelnen Bestandteile
cd temp
call saxon -o xdiml_front.xml front.xml %home%xsl\ooo_2xdimlfront.xsl name="%1"
scale="%SCALE%"
for %%i in (chapter?.xml) do saxon -o xdiml_%%i %%i
%home%xsl\ooo_2xdimlbody.xsl name="%1" scale="%SCALE%"
call saxon -o xdiml_back.xml back.xml %home%xsl\ooo_2xdimlback.xsl name="%1"
scale="%SCALE%"
cd ..
REM Zusammenfügen der einzelnen Bestandteile zu einem xdiml file
call perl %home%pl\ooo_completeLit.pl "%1" "%BIBLIO%"
```

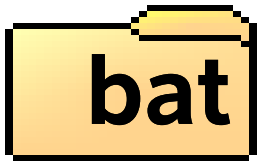
Organisation



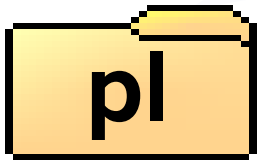
OpenOffice (StarOffice)
XMLSpy (XML-Editor)
Java
Perl
Saxon
Winzip
Internet Browser
Texteditor
Graphic program
Converter

Save as XDiML II

Converter



- ooo2dimlstyles.bat
- ooo2html.bat
- ooo2xdiml.bat
- ooo2xdimlA.bat
- ooo2xdimlparts.bat
- ooo2xdimlpartsA.bat
- saxon.exe



- ooo_bibliography.pl
- ooo_bibliographyA.pl
- ooo_complete.pl
- ooo_completeLit.PL
- ooo_doctype.pl
- ooo_entities.pl
- ooo_zeilen.pl

- ooo_2dimlstyles.xsl
- ooo_2dimlstylesfilter.xsl
- ooo_2errorfilter.xsl
- ooo_2parts.xsl
- ooo_2xdiml.xsl
- ooo_2xdimlA1.xsl
- ooo_2xdimlA2.xsl
- ooo_2xdimlback.xsl
- ooo_2xdimlbody.xsl
- ooo_2xdimlfront.xsl
- ooo_2xdimlfürAbschnitte.xsl
- ooo_body.xsl
- ooo_common.xsl
- ooo_error.xsl
- ooo_font.xsl
- ooo_frontback.xsl
- ooo_replace.xsl
- ooo_styles.xsl
- ooo_tables.xsl

Summary

- Theses, created with OpenOffice and the template dissertation-hu.stw are qualified for long term preservation in XML, in particular in XDiML.
- The conversion process from the OpenOffice document to XDiML runs mainly automatical, but will never be completely automatical, because every document has its specific individualities.
- The Open Office file format is only partly an adequate format for transforming into XDiML, for instance because of it's lack for nesting of chapters.