Save as XDiML (DissertationMarkupLanguage),

Writing and Converting digital Theses and Dissertations using OpenOffice.org

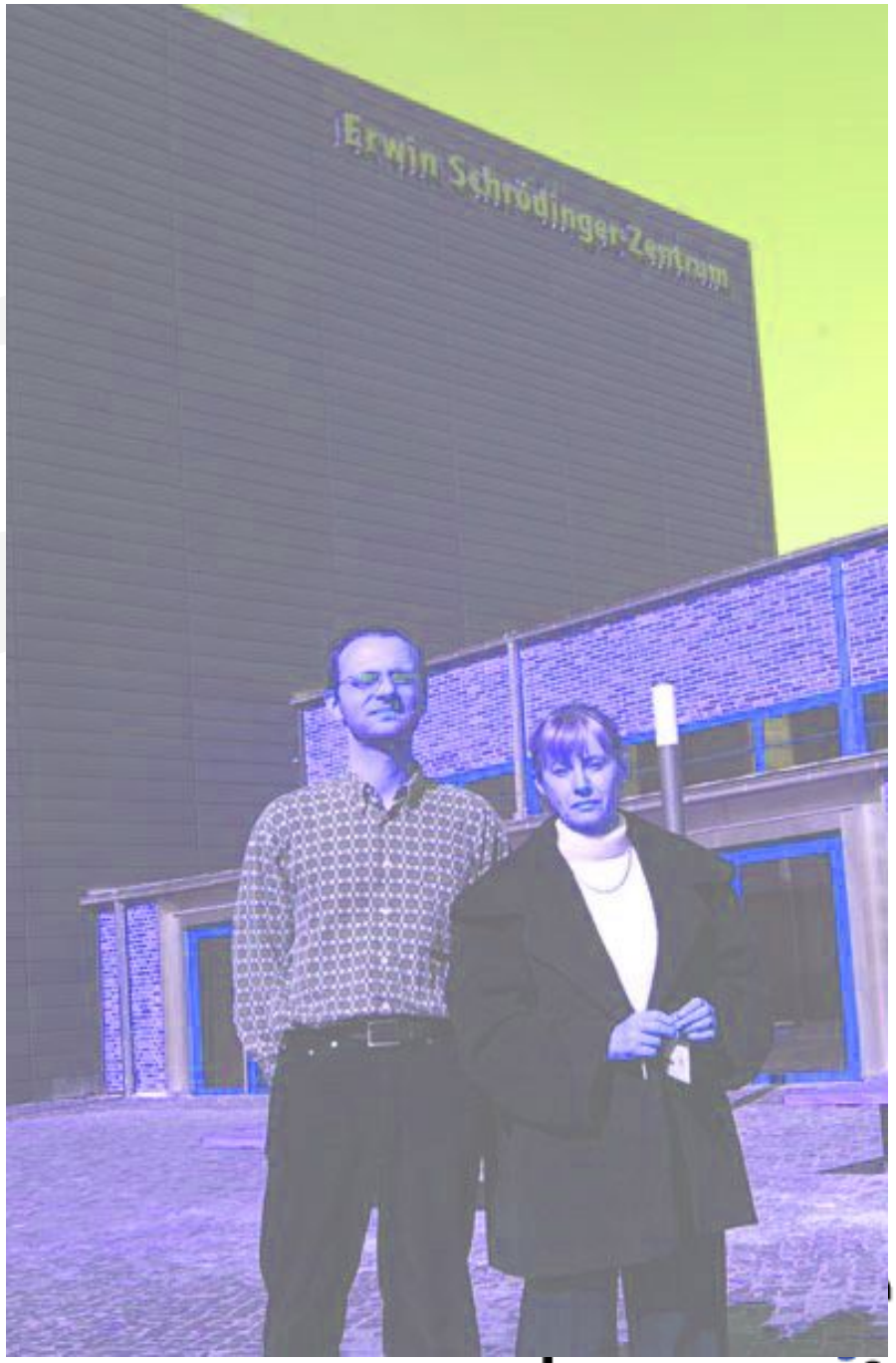
by
Sabine Henneberger and
Matthias Schulz
edoc@cms.hu-berlin.de

Agenda

- About the Speakers
- Part 1.
  - Aspects of Electronic Publishing
  - What is XDiML?
- Part 2.
  - Converting and Saving to XDiml-Format
    - Converting from Word-X
    - The “Save as Xdiml”-Filter
  - Writing your Theses in OpenOffice.org
    - The “Dissertation”-Menu
- Part 3.
  - Demonstration
- Conclusions
- Links

About the 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 main parts of OpenOffice support for XDiML
- Jakob Voss and Matthias Schulz developed the XdiML DTD (XML-Version of DiML DTD)



Expectations to Electronic Publishing

- View of authors
  - create and edit; publishing / dissemination, intellectual property rights / authenticity / integrity, question of time,...
- View of users
  - availability, retrieval capabilities, authenticity, ...
- View of libraries
  - acquisition, exploitation, cataloguing, long term archiving, authenticity, ...
- View of computing centers
  - availability, bandwidth of computer network, retrieval, long term archiving, searching machines, storage capacity,...
- View of publishers
  - technological process, quality control, dissemination, marketing, …

Changes in Publishing Workflows

- The main points of the publishing workflow
  - Creation
  - Archiving
  - Retrieval
  - And **Problems**

Creating an Electronic Document

- Demands of the authors
  - Modern text processor
  - Supporting tools for multimedia applications
  - Guarantee of integrity and authenticity of my document
  - Long term archiving
  - Short publication times
  - Worldwide availability
- Requirements to the authors
  - Do not use proprietary systems or file formats
  - Use standards or at least common rules
  - In order to support the retrieval create a structured text
  - Do not use your own system of citation

Wordprocessors and DTP

- Word 76,4%
- LaTeX 21,0%
- Corel WordPerfect 1,1%
- FrameMaker 0,9%
- OpenOffice.org / Staroffice 0,6%
- Submission of Xdiml 0,0%

- 727 dissertation and doctor theses (1997-2003) available at Humboldt University

Archiving

- Long term preservation for 10 ... years
- Using standardized document formats ISO8879
- Easy reconversion or transformation into new presentation or print formats
- Including of multimedia objects

Retrieval

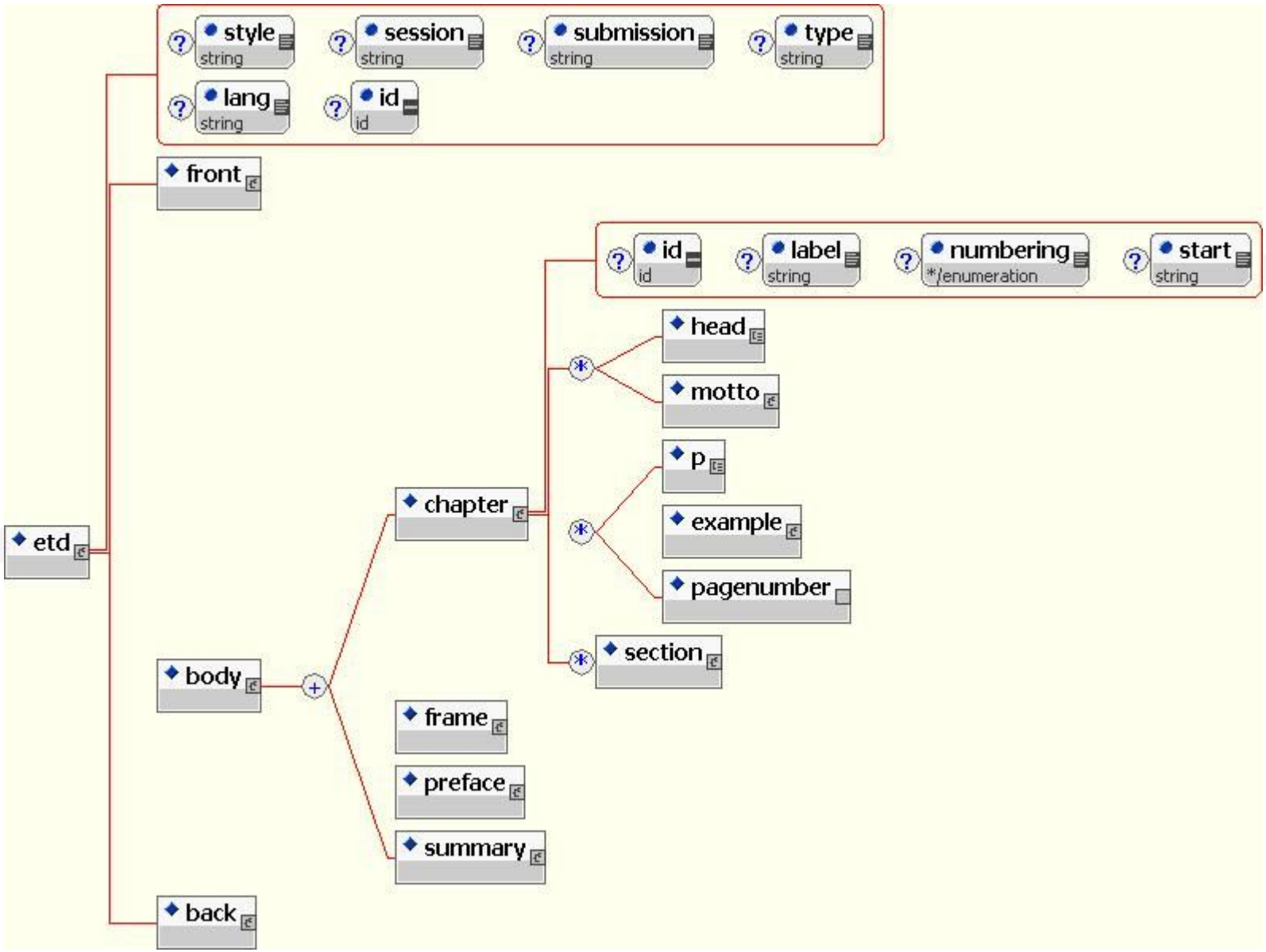
- Using document structure and semantic tags
- Detailed search
- Automated cataloging
- Information extraction (e.g. citation index)
- Value of highly structured information

What is XDiML?

- XDiML (DissertationMarkupLanguage in XML)
- First DTD, the DiML.dtd in SGML out of the ETD-ML.dtd of Virginia Tech in 1997
- ETD-ML was developed from Yuri Rubinsky(SGML Pioneer) and Neil Kipp

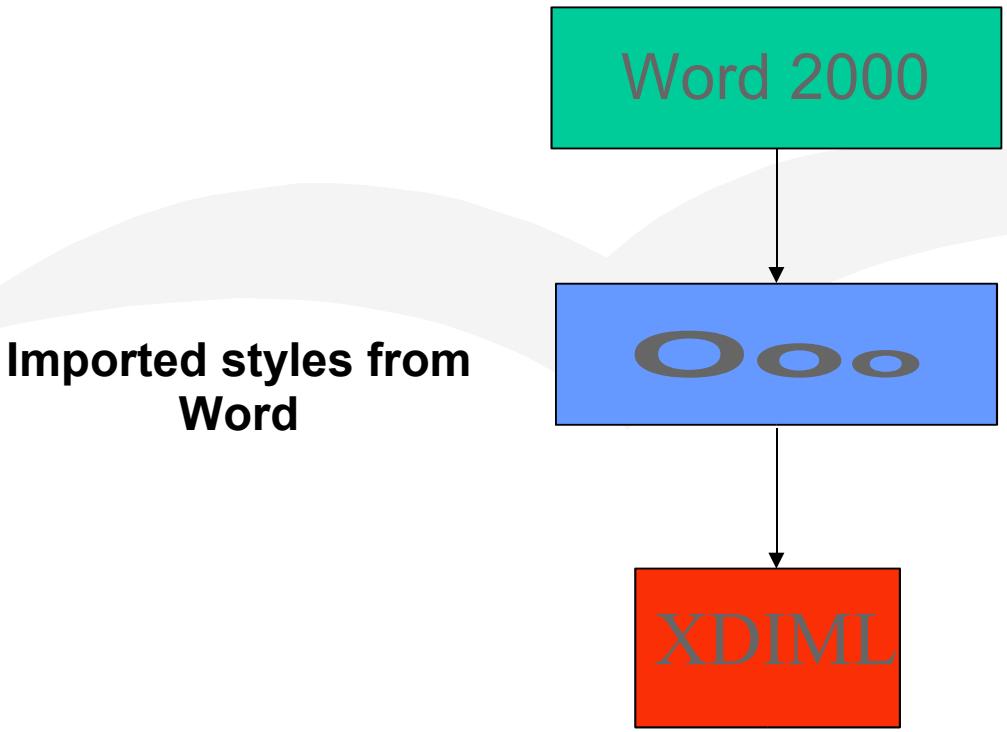
- The DTD has a document structure like books.
  - Root Element: etd, Childs: front, body, back and then chapter...
- The structure of DTD is modulbased like TEI-DTD
  - Moduls for MathML and other special DTDs
  - And for the main parts of the DTD (chapter...)

View into XDiML



Save as XDiML

From Word 2000 to XDiML

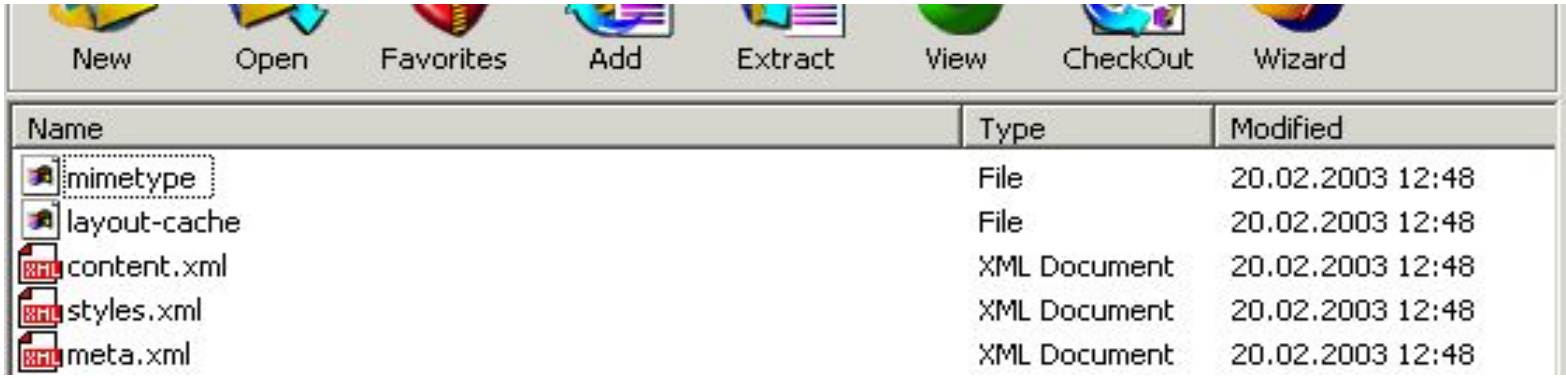


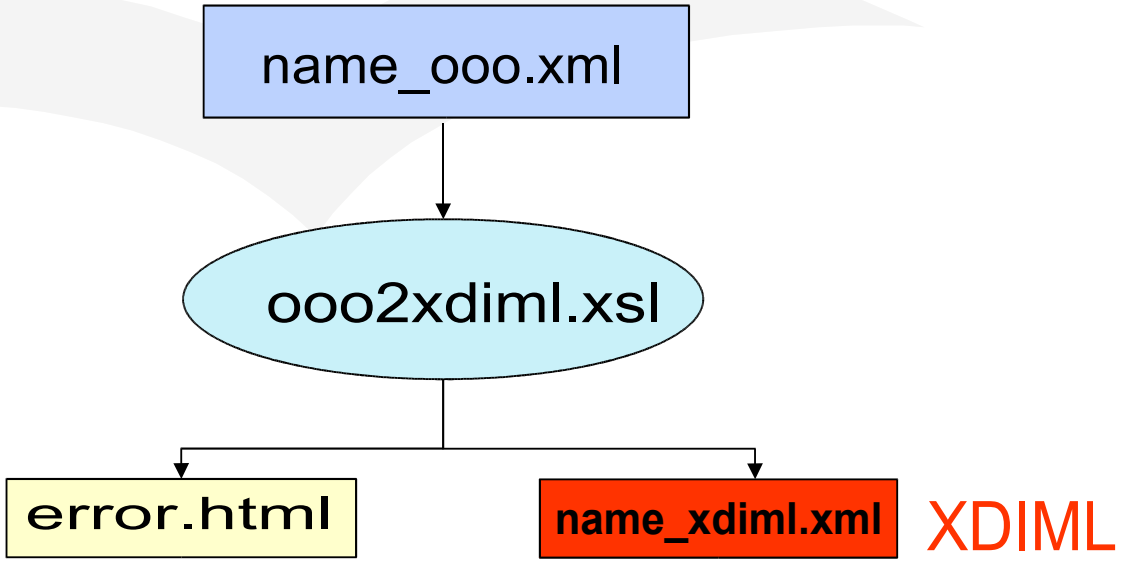
Save as XDiML

1st approach:

name.sxw

content.xml
renamed to
name_ooo.xml





Stucture of name_ooo.xml (Overview)

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE office:document-content PUBLIC "-//OpenOffice.org//DTD OfficeDocument 1.0//EN" "C:\Office DTD\office.dtd">
<office:document-content xmlns:office="http://openoffice.org/2000/office" xmlns:style="http://openoffice.org/2000/style" ...>
<office:automatic-styles>..... </office:automatic-styles>
...
<office:body>..... </office:body>
</office:document-content>
```

Problems to solve: (A) Assigning the tags

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"/>
```

Problems to solve (A)

```
<office:automatic-styles>
<style:style style:family="paragraph" style:name="P1" style:parent-style-name =“Titel">
<style:properties fo:margin-left="0cm" fo:margin-right="0.635cm" fo:text-indent="0cm" style:auto-text-indent="false"/>
</style:style>
```

Problems to solve: (T) Transforming the structure

Chapters and subchapters as an example

<office:body>

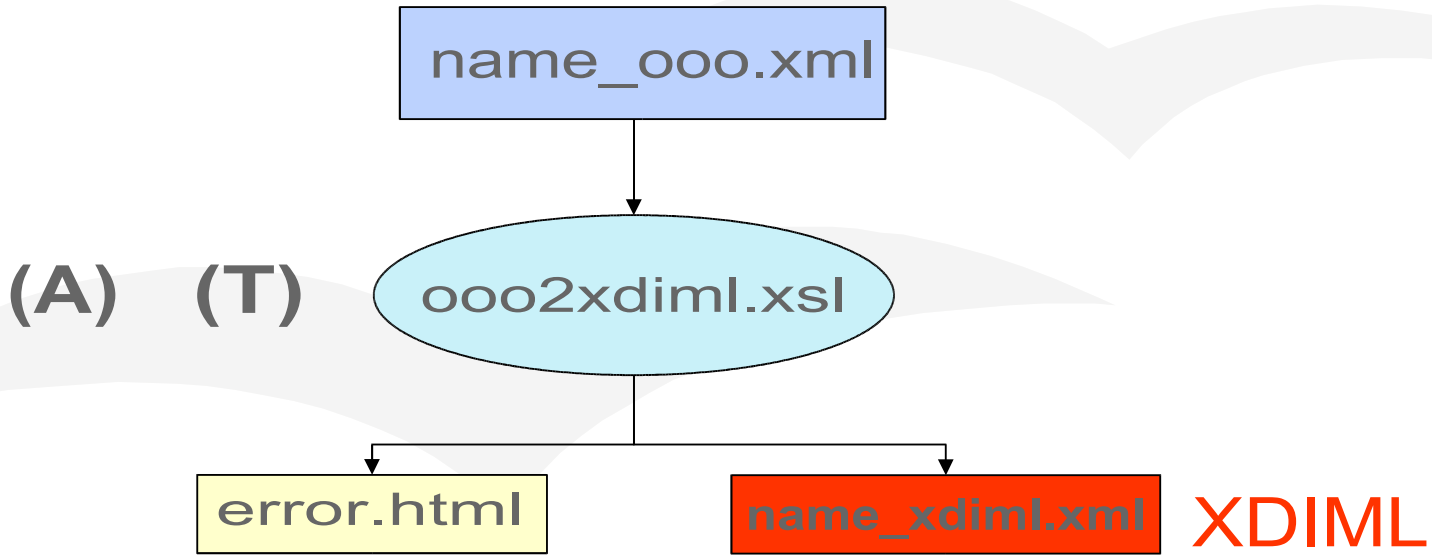
```
<text:h text:style-name="P20" text:level="1">Heading 1</text:h>

<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>
```

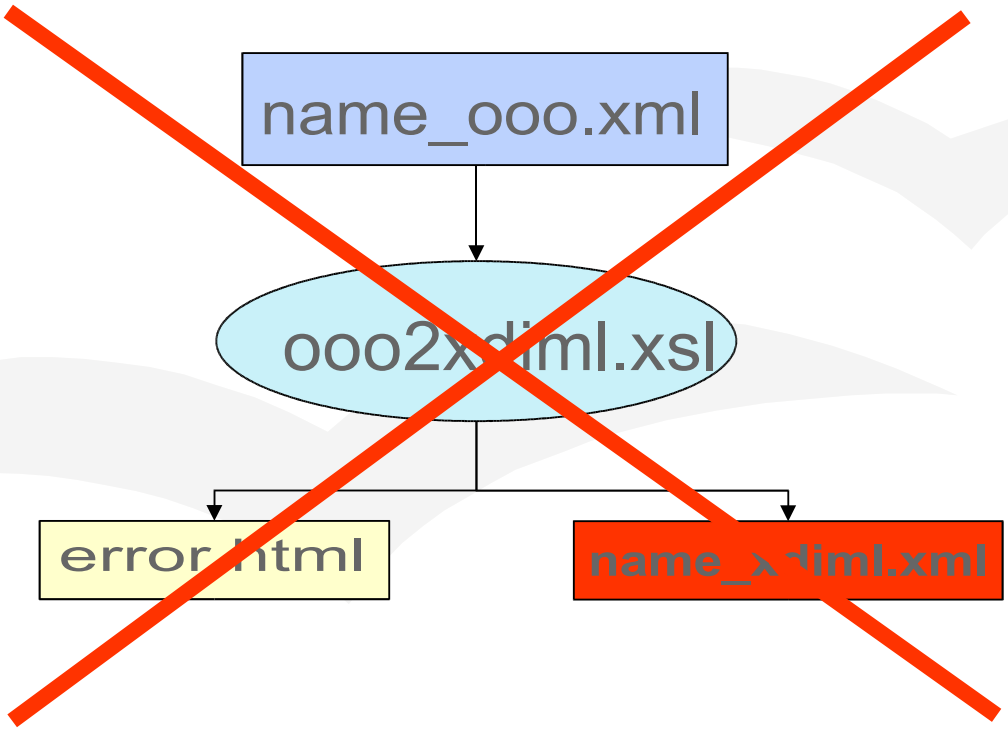
Save as XDiML

1st approach again:



Save as XDiML

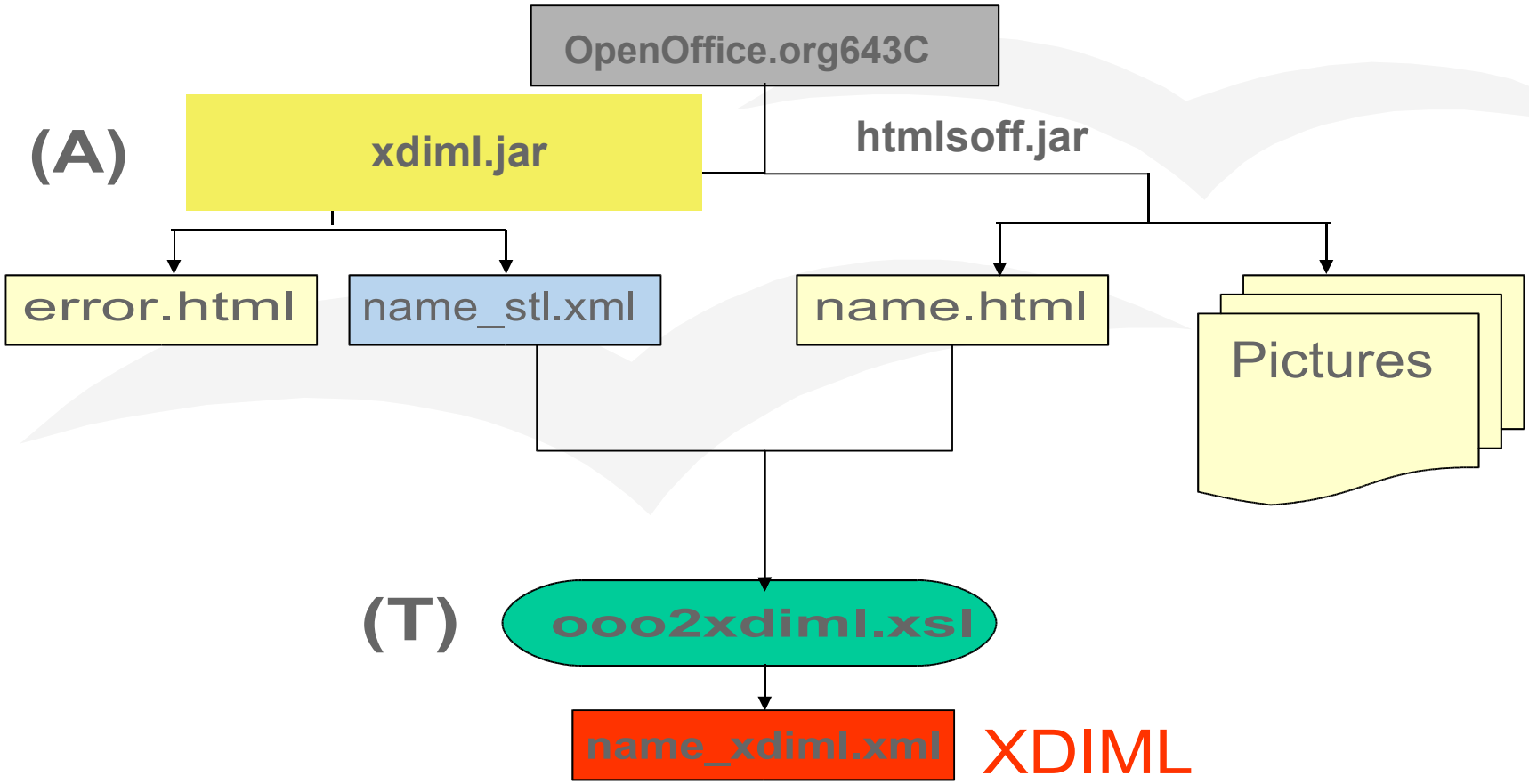
Problems of processing



TIME !!!

Save as XDiML

2nd approach



2nd approach problems

Filter xdiml.jar:

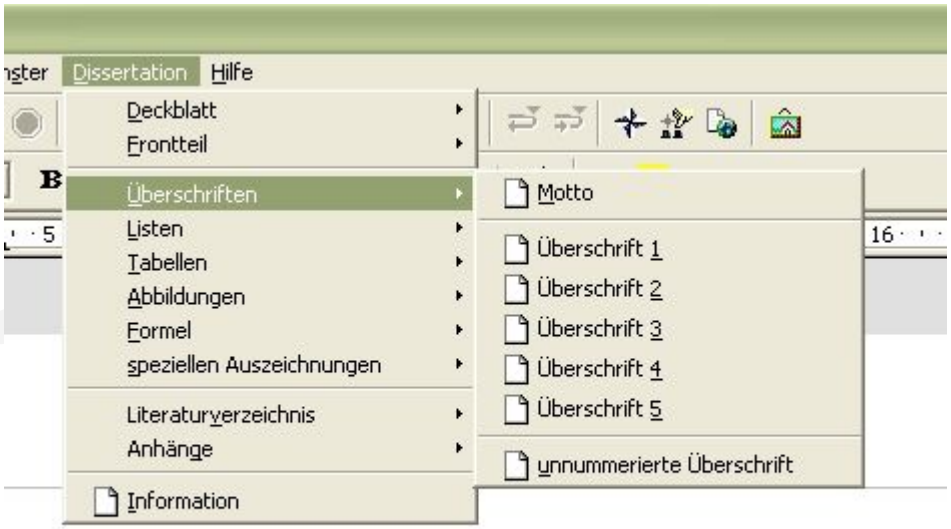
- Error.html inside of name_stl.xml
- Process pictures directly?
- In case of OLE no output
- Performance (time) not satisfying
- Large documents?
- Working with OO.org 1?

Stylesheet ooo2xdiml.xsl:

- Time (tables), document with number of table-cells>3000 had to be separated in 2 parts
- Document with 2400 table cells 17 min

Write and Save your Theses

- What does the student or doctoral student need for writing?
  - Template for dissertation
  - Manual, support and lesson concept for OpenOffice.org
  - Web Page



The Template “Dissertation”

- Main aspects are:
  - “Formatting” the specific content (Document type)
  - Semantic markup
  - Excact and better transformation to XdiML Format
- Second aspect (student's aspect)
  - Help the authors to format the dissertation!

Part 3: Live Demonstration

Conclusions

- For special document types we need filters and templates for OpenOffice.org.
- The user needs help: tools (menu), lessons, manuals and support.
- The conversion is not an one way process. It is a very complex process with different tools or even a process with different parts and file formats.

Links

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