preserving our Digital Assets

PLANETS, Document Conversion Tools and the OpenXML/ODF Translator

Document Interoperability Initiative

Brussels, 12 November 2009

Wolfgang Keber (wk@dialogika.de)

Overview

- PLANETS Project
- Document conversion tools
  - Objectives
  - Technical approach
  - Demo
- OpenXML/ODF Translator
  - Overview
  - Next steps (ISO 29500 compatibility)

preserving our Digital Assets

PLANETS

Preservation and

Long-term

Access through

Networked

Services

PLANETS is a four-year project co-funded by the European Union under the Sixth Framework Programme to address core digital preservation challenges.

“The primary goal for Planets is to build practical services and tools to help ensure long-term access to our digital cultural and scientific assets. “

(excerpt from http://www.planets-project.eu/)

preserving our Digital Assets

Partners

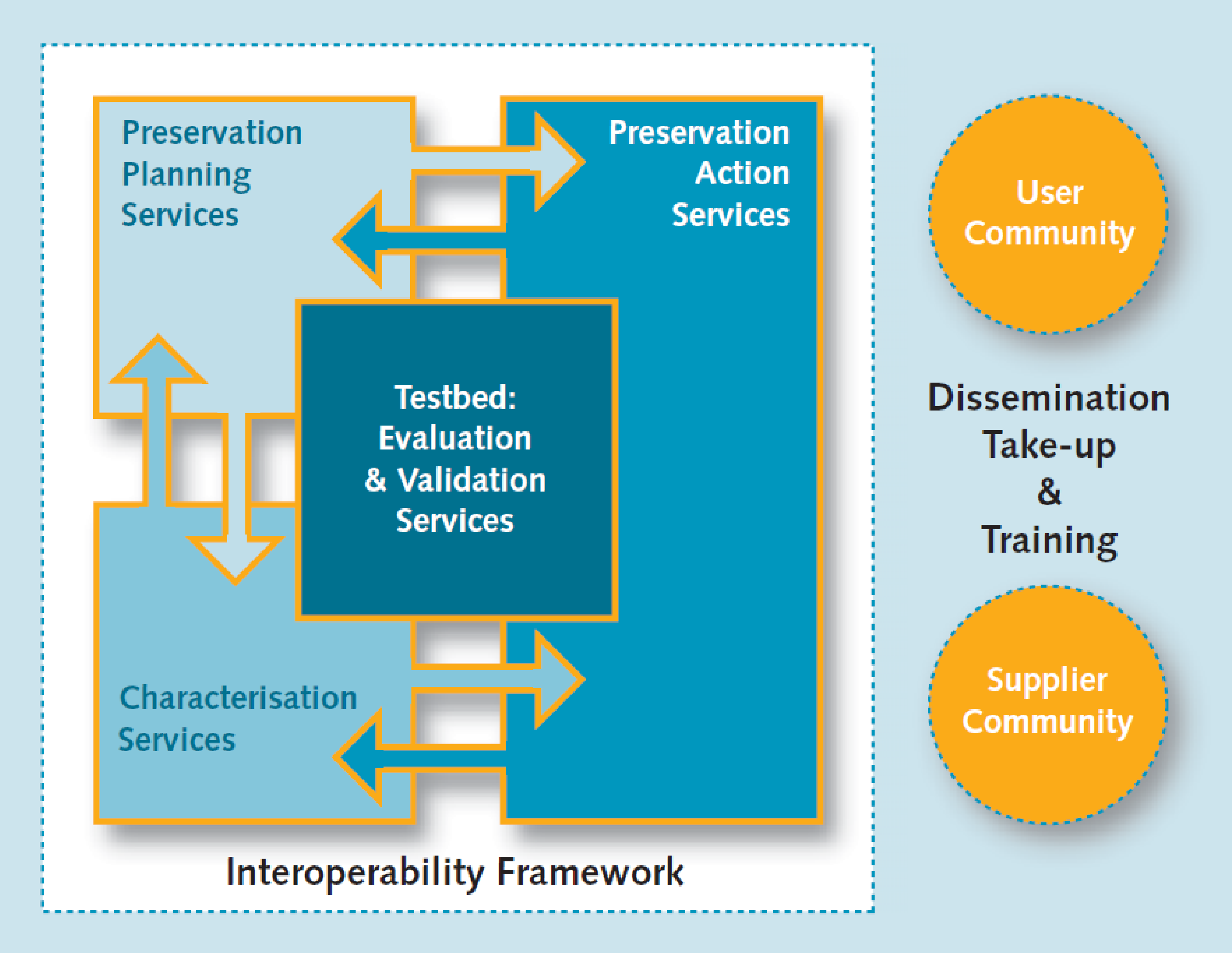
- The British Library
- National Library, Netherlands
- Austrian National Library
- State and University Library, Denmark
- Royal Library, Denmark

- National Archives, UK
- Swiss Federal Archives
- National Archives, Netherlands

- Hatii at University of Glasgow
- University of Freiburg
- Technical University of Vienna
- University at Cologne

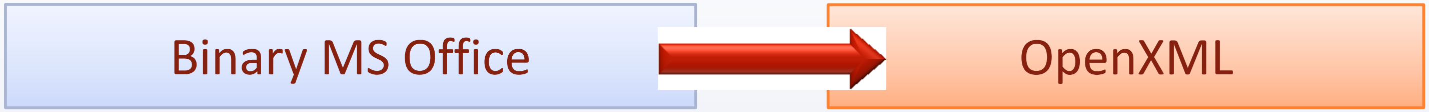
- Tessella Plc
- IBM Netherlands
- Microsoft Research, Cambridge (with DIaLOGIKa as a Microsoft partner)
- ARC Seibersdorf research

Sub Projects

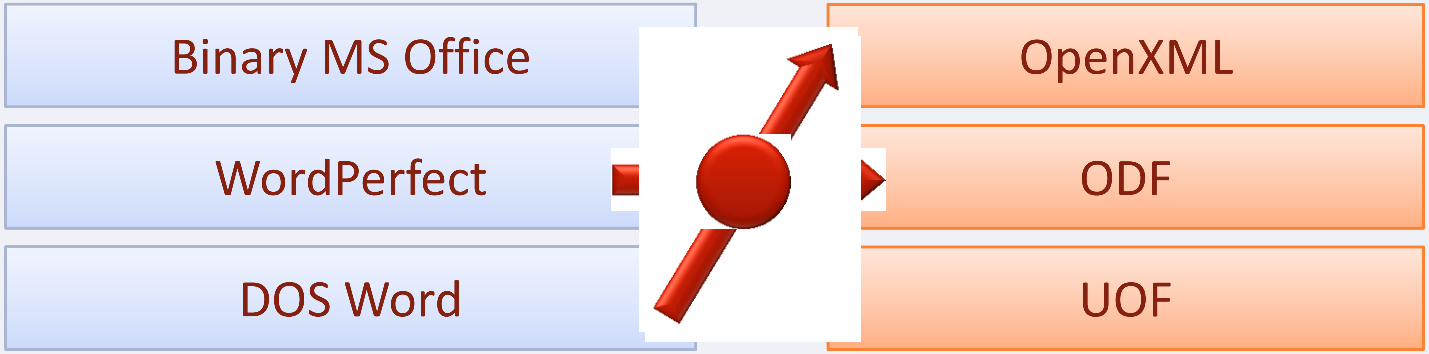

Preservation Planning Services
Preservation Action Services
Testbed: Evaluation & Validation Services
Characterisation Services
Interoperability Framework
User Community
Dissemination Take-up & Training
Supplier Community

Microsoft & PLANETS: Office Documents

- Microsoft Research's role within PLANETS:
  - Conversion of binary Microsoft Office Documents into Office Open XML File Format (OpenXML)



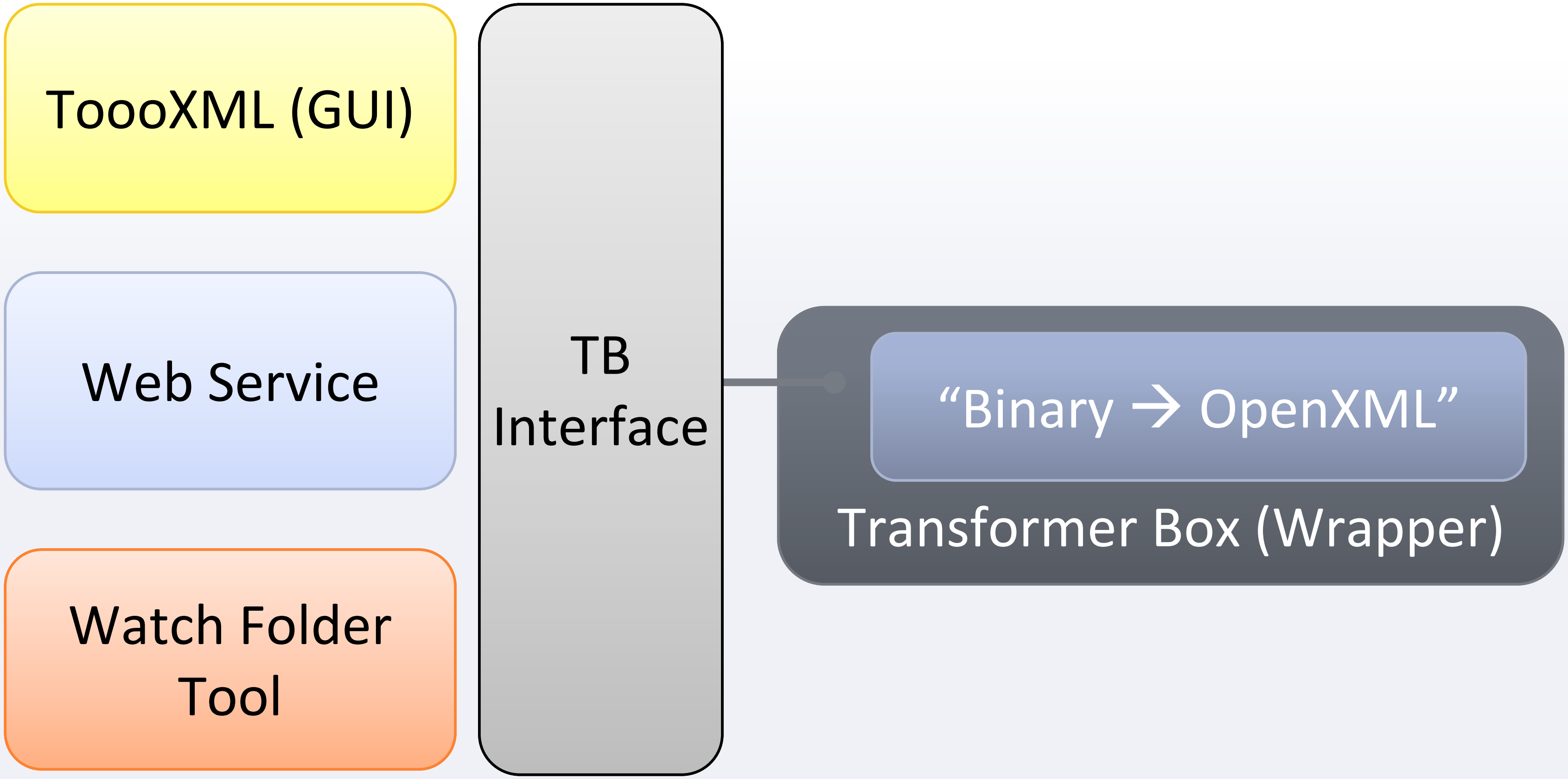
- We extended the effort to include other formats
  - More legacy formats, e.g. WordPerfect
  - Other open standards, e.g. Open Document Format.



Document Conversion Tools – Our Approach

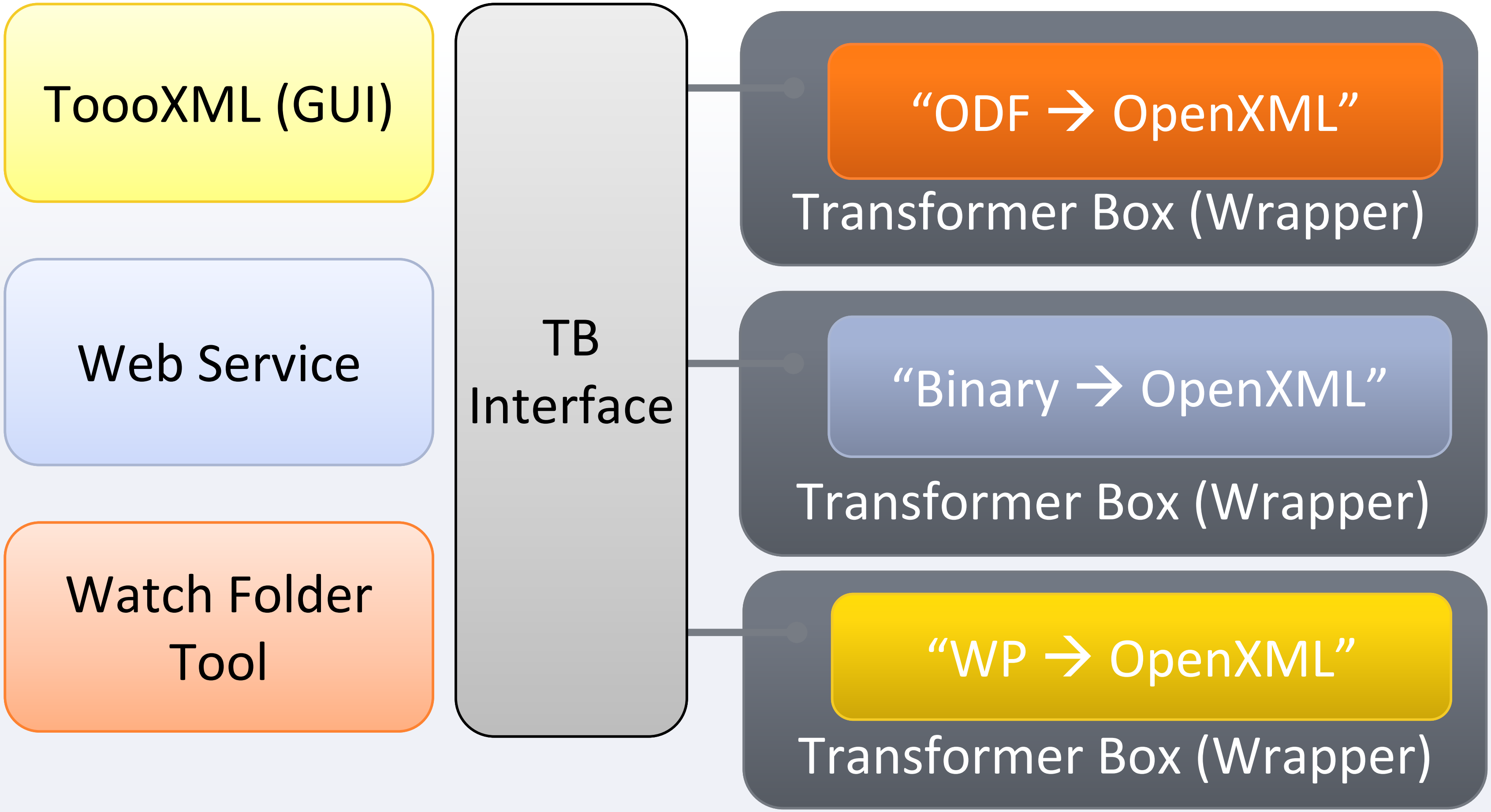
- Three-step approach, resulting in a modular and extendible infrastructure
  - Identify existing conversion tools and libraries
  - Wrap these tools and libraries into re-usable components
  - Integrate these components into PLANETS and other systems.
- If possible, do not use the office applications (e.g., Microsoft Office or OpenOffice.org)
  - They are designed as interactive applications
  - Message boxes might pop up ("Do you want …")
  - Unclear license question when running on a server.

Reusable Components



preserving our Digital Assets

Extensible Architecture



ToooXML (GUI)

Web Service

Watch Folder Tool

TB Interface

“ODF → OpenXML”

Transformer Box (Wrapper)

“Binary → OpenXML”

Transformer Box (Wrapper)

“WP → OpenXML”

Transformer Box (Wrapper)

More Technical Details

- Currently two types of wrappers for
  - Command-line tools (stand-alone executables)
    - OpenXML/ODF Translator (OpenXML ⇔ ODF)
    - OpenXML Document Viewer (OpenXML ⇨ HTML)
  - Microsoft conversion libraries (CNV libraries)
    - WordPerfect ⇨ RTF
    - RTF ⇨ OpenXML
    - …
- Wrappers can be chained
  - WordPerfect ⇨ RTF ⇨ OpenXML ⇨ ODF.

Supported Formats

- Source formats
  - WordPerfect 5
  - WordPerfect 6
  - DOS Word
  - Word 2, 6, 95
  - Office 97-2003
  - RTF
  - ODF
  - OpenXML

- Target formats
  - OpenXML
  - ODF
  - UOF
  - HTML
  - XCDL (format defined in PLANETS/PC)

preserving our Digital Assets

ToOOXML++ Demo

- ToOOXML++
  - UI for demonstrating the conversion tools
  - Allows documents to be selected
  - Automatically determines the document format
  - Offers a list of available target formats.
- Virtual PC
  - Pre-installed ToOOXML++
  - Legacy applications to view documents in their native applications
    - WinWord 2
    - WordPerfect 5
    - …

OpenXML/ODF Translator (1)

- Open Source project hosted on SourceForge (http://sourceforge.net/projects/odf-converter)
- Developed under a liberate BSD-like license
- Several companies involved
  - Sonata and DIaLOGIKa (development & testing)
  - Novell (OpenOffice.org/Linux integration)
  - Microsoft (funding and coordination)

OpenXML/ODF Translator (2)

- Available in three variants
  - Add-in for Office 2000, XP and 2003
  - Add-in for Office 2007
  - Command-line tool (Office apps not required)
- All use the same translation kernel based on XSLT-technology
  - Pre- and postprocessing for special purposes
  - .Net Framework 2.0/C# for Office integration
- Compatibility with other platforms via Mono (e.g. Linux)
- Test suites based on
  - Documents containing specific features
  - Real documents found in the Internet and from other sources (public administration)

Next Major Step – ISO 29500 Compatibility (1)

- Office Open XML “standards”
  - Starting point ECMA-376 1st Edition (December 2006)
    - Office 2007 & File Format Compatibility Pack
    - Current OpenXML/ODF Translator
  - Evolved to ISO/IEC 29500:2008
  - Identical with ECMA-376 2nd Edition (December 2008)
    - Office 2010 (more in Doug’s presentation)
    - Next major release of OpenXML/ODF Translator
    - File Format Compatibility Pack?

preserving our Digital Assets

preserving our Digital Assets

Next Major Step – ISO 29500 Compatibility (2)

- It’s a rugged way to ISO 29500
  - “transitional” vs. “strict”
  - “producer” vs. “consumer”
  - How can we test the translator?
    - How can we validate the created ISO 29500 documents?
    - How can we create ISO 29500 test documents?
  - What has actually been changed?
    - There is no nice comprehensive Excel sheet with all changes
    - Schema comparison
    - Responses from NBs
    - Resolutions from BRM
    - Final standard

Example: BRM Resolution 7

- *The BRM resolves to accept the editing instructions contained in http:/.../Response_DE-0028_dates_v9.doc in replacement of R 18 and R 43, but with the following corrections: the words "ISO value" shall be replaced by the words "ISO 8601 value"; page 9, line 24 shall be restored and line 23 shall be marked "transitional"; and all changes as well as choices among alternatives rendered necessary by the Decision above shall be made—results of the vote: 19 in favour; 3 against (EC, US, ZA); 9 abstentions (JP, KR, IE, AU, BR, CN, NL, MX and GR): so resolved*
- *R 18 and R 43*
- *Response_DE-0028_dates_v9.doc*

preserving our Digital Assets

Q & A

Document Interoperability Initiative

Brussels, 12 November 2009

Wolfgang Keber (wk@dialogika.de)

DIaLOGIKa