MIXED
MIGRATION TO INTERMEDIATE XML FOR ELECTRONIC DATA

MIXED: repository of durable file format conversions

Dirk Roorda and René van Horik

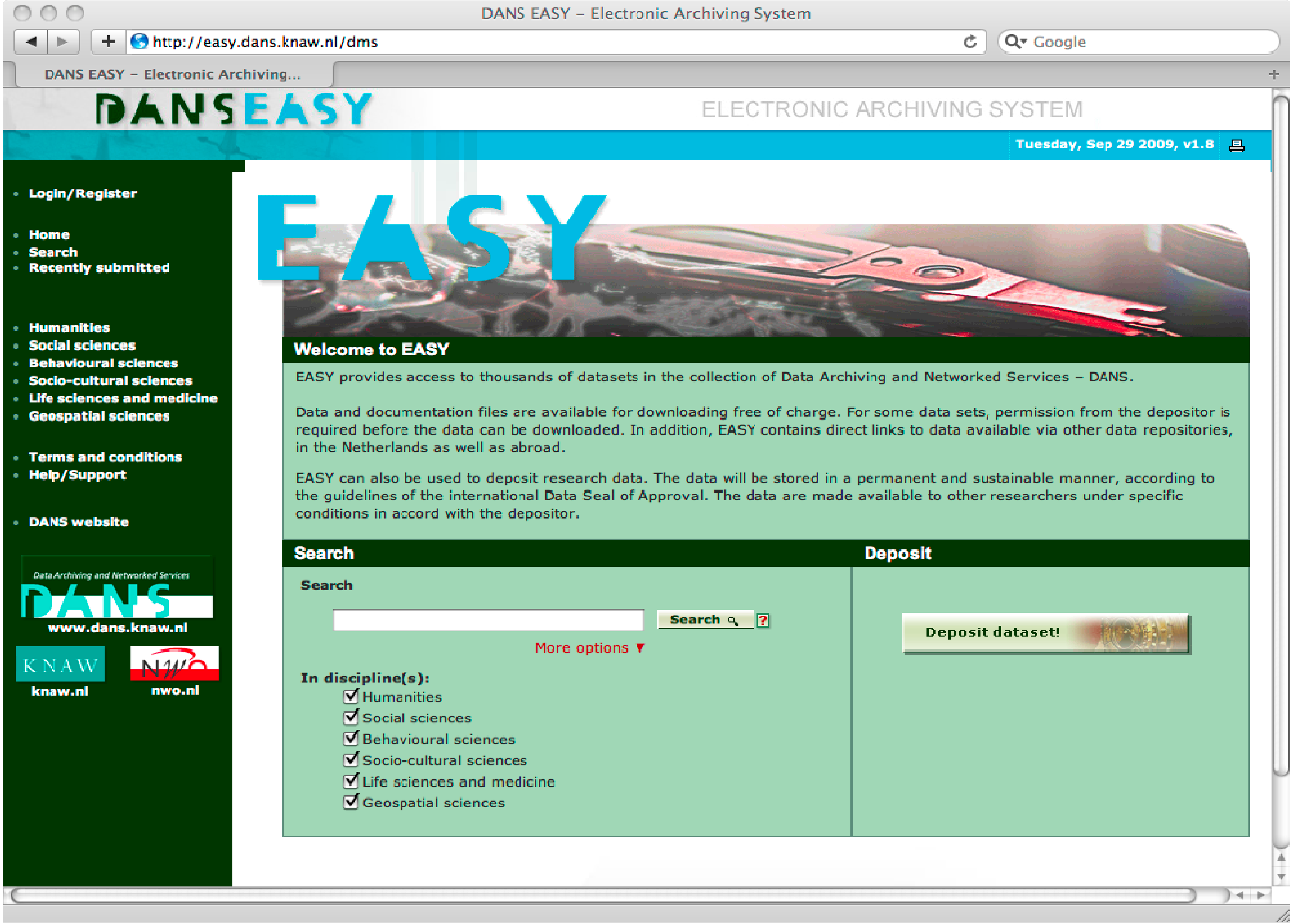
Data Archiving and Networked Services
DANS

MIGRATION TO INTERMEDIATE XML FOR ELECTRONIC DATA

The aim of the MIXED project is to develop a sound theoretical framework for the curation of file formats and practical services and tools that support this framework

Outline

- Scientific data-archives and durability: the DANS case
- MIXED: a tool to keep data formats usable
- The “Smart migration” strategy
- SDFP: Standard Data Formats for Preservation
- MIXED Software: Framework and Plugins
- Discussion



Welcome to EASY

EASY provides access to thousands of datasets in the collection of Data Archiving and Networked Services – DANS.

Data and documentation files are available for downloading free of charge. For some data sets, permission from the depositor is required before the data can be downloaded. In addition, EASY contains direct links to data available via other data repositories, in the Netherlands as well as abroad.

EASY can also be used to deposit research data. The data will be stored in a permanent and sustainable manner, according to the guidelines of the international Data Seal of Approval. The data are made available to other researchers under specific conditions in accord with the depositor.

In discipline(s):

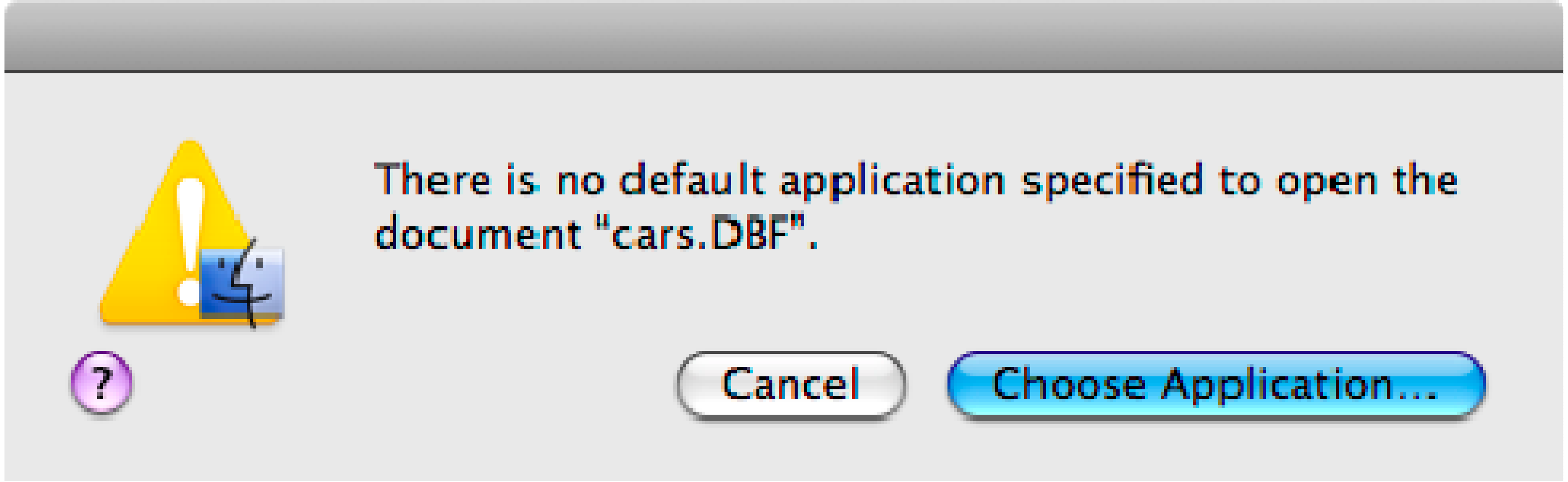
- ☑ Humanities
- ☑ Social sciences
- ☑ Behavioural sciences
- ☑ Socio-cultural sciences
- ☑ Life sciences and medicine
- ☑ Geospatial sciences

Number of files in DANS archive: 495.437 (august 2009)

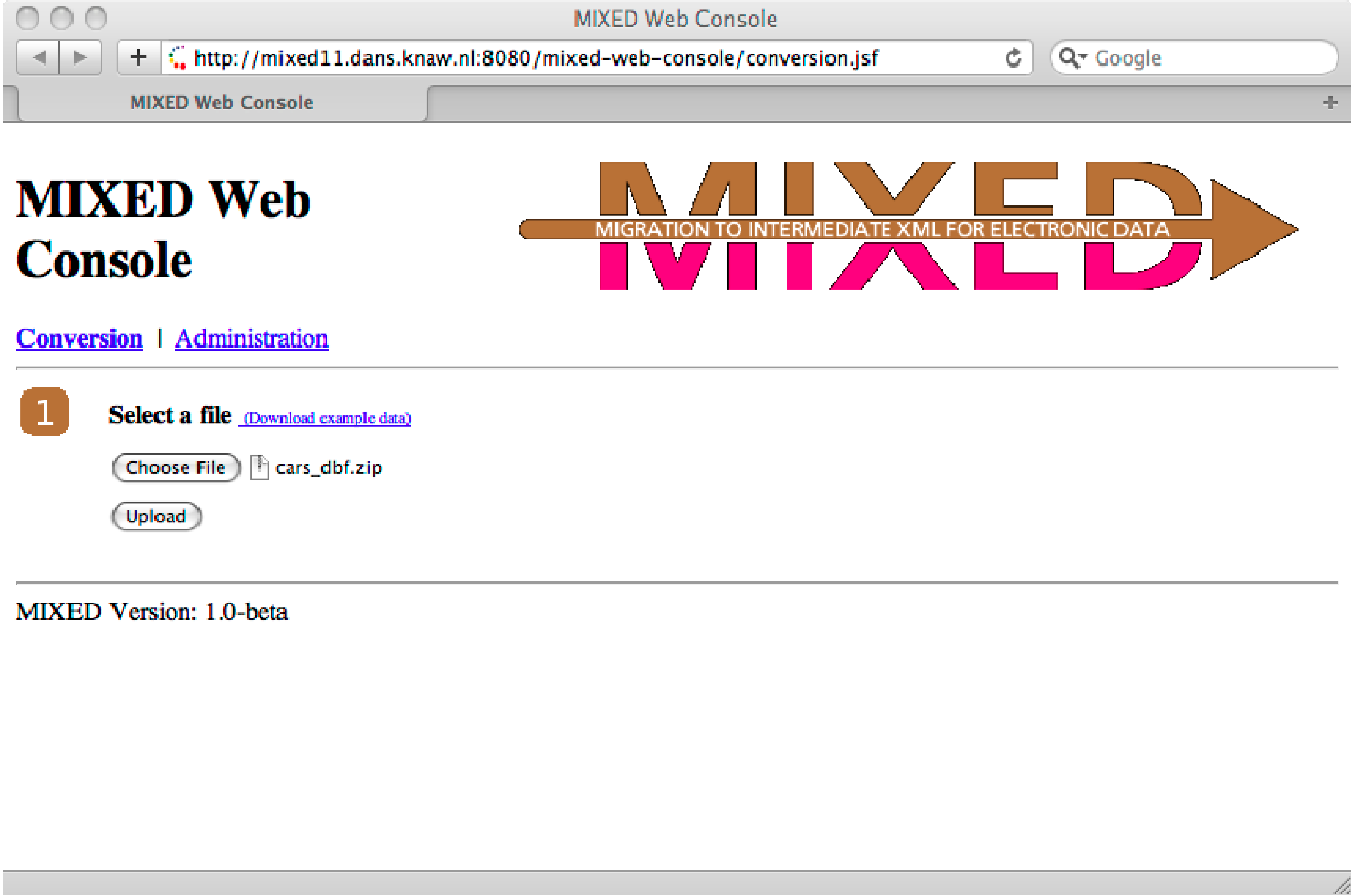
Diversity of formats

JPEG
TIFF
PDF
ASCII
WORD
HTML
Postscript
AutoCad
WordPerfect
XML
MS Access
Powerpoint
Etc

DataPerfect
Lotus 123
CSV
Filemaker Pro
Paradox
WordStar
Harvard Graphics
Quatro pro
SPSS
DTD
SAS
Java Script
Etc



There is no default application specified to open the document "cars.DBF".



MIXED Web Console
http://mixed11.dans.knaw.nl:8080/mixed-web-console/conversion.jsf
Google
MIXED Web Console
MIXED
MIGRATION TO INTERMEDIATE XML FOR ELECTRONIC DATA
Choose File
Upload

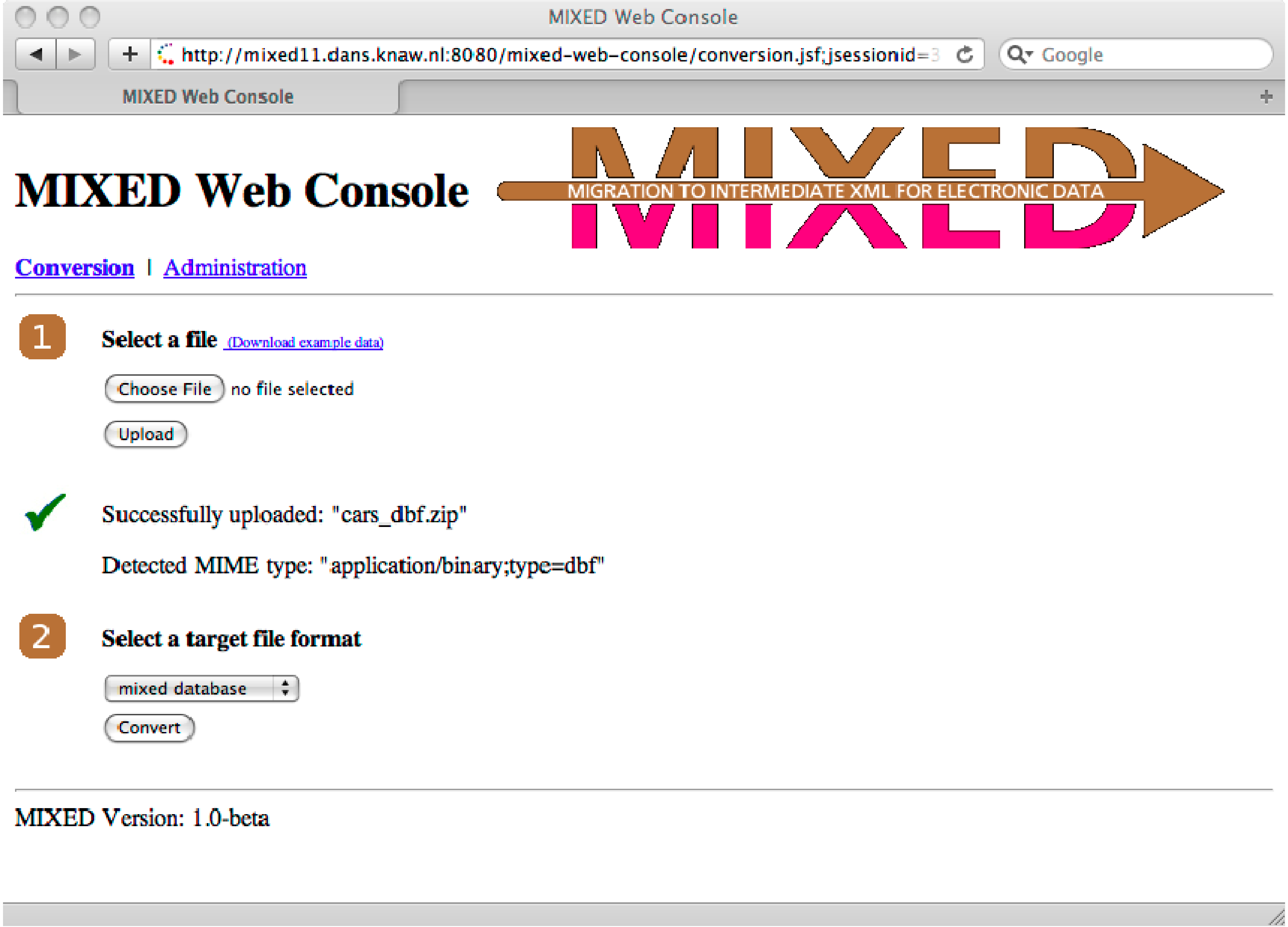
MIXED Web Console

Conversion | Administration

1 Select a file (Download example data)

cars_dbf.zip

MIXED Version: 1.0-beta



http://mixed11.dans.knaw.nl:8080/mixed-web-console/conversion.jsf;jsessionid=3

MIXED Web Console

Conversion | Administration

1 **Select a file** (Download example data)

no file selected

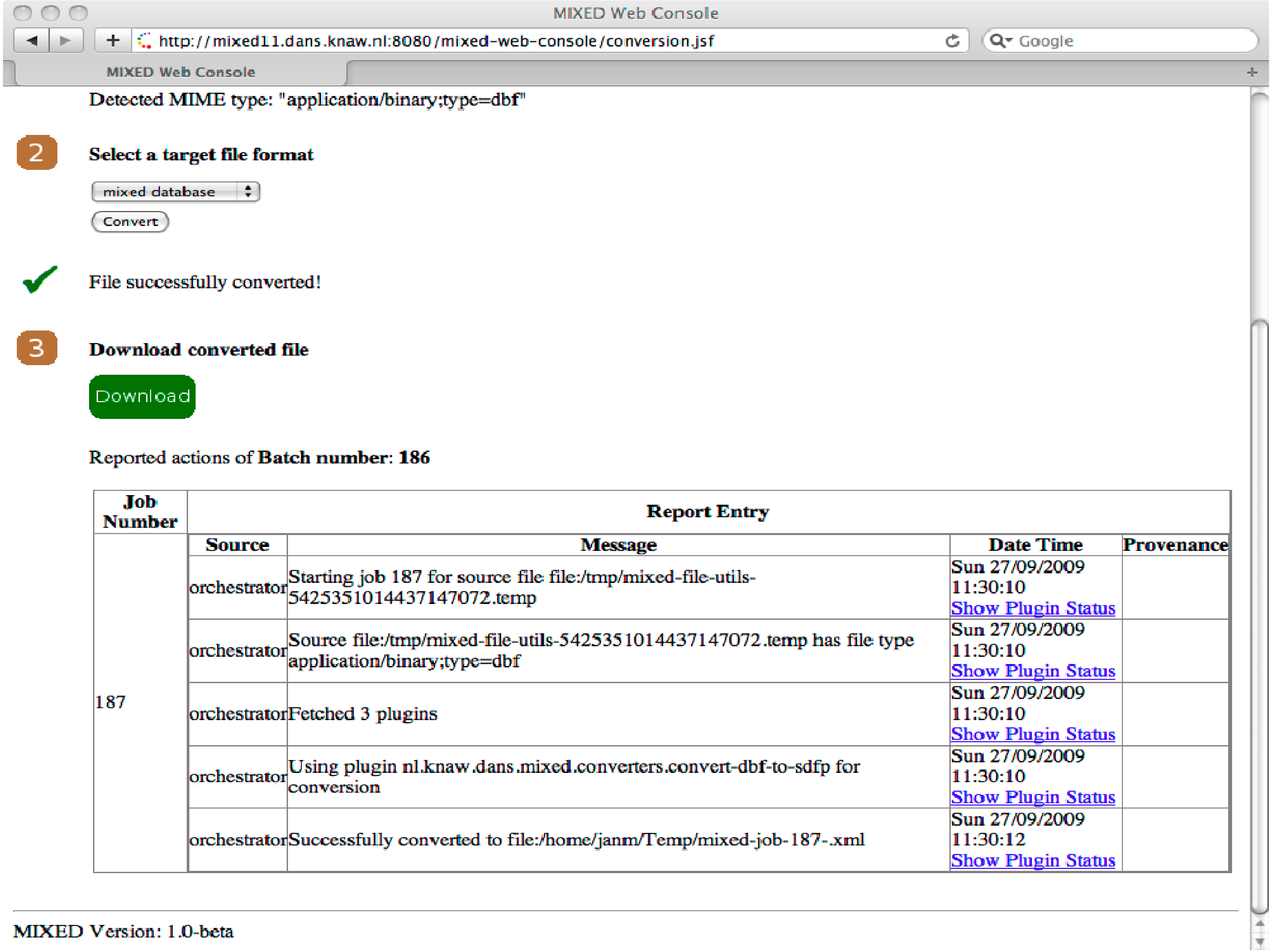
✓ Successfully uploaded: "cars_dbf.zip"

Detected MIME type: "application/binary;type=dbf"

2 **Select a target file format**

mixed database

MIXED Version: 1.0-beta


MIXED Web Console
http://mixed11.dans.knaw.nl:8080/mixed-web-console/conversion.jsf
Google
MIXED Web Console

Detected MIME type: "application/binary;type=dbf"

2 **Select a target file format**

mixed database

Convert

✓ File successfully converted!

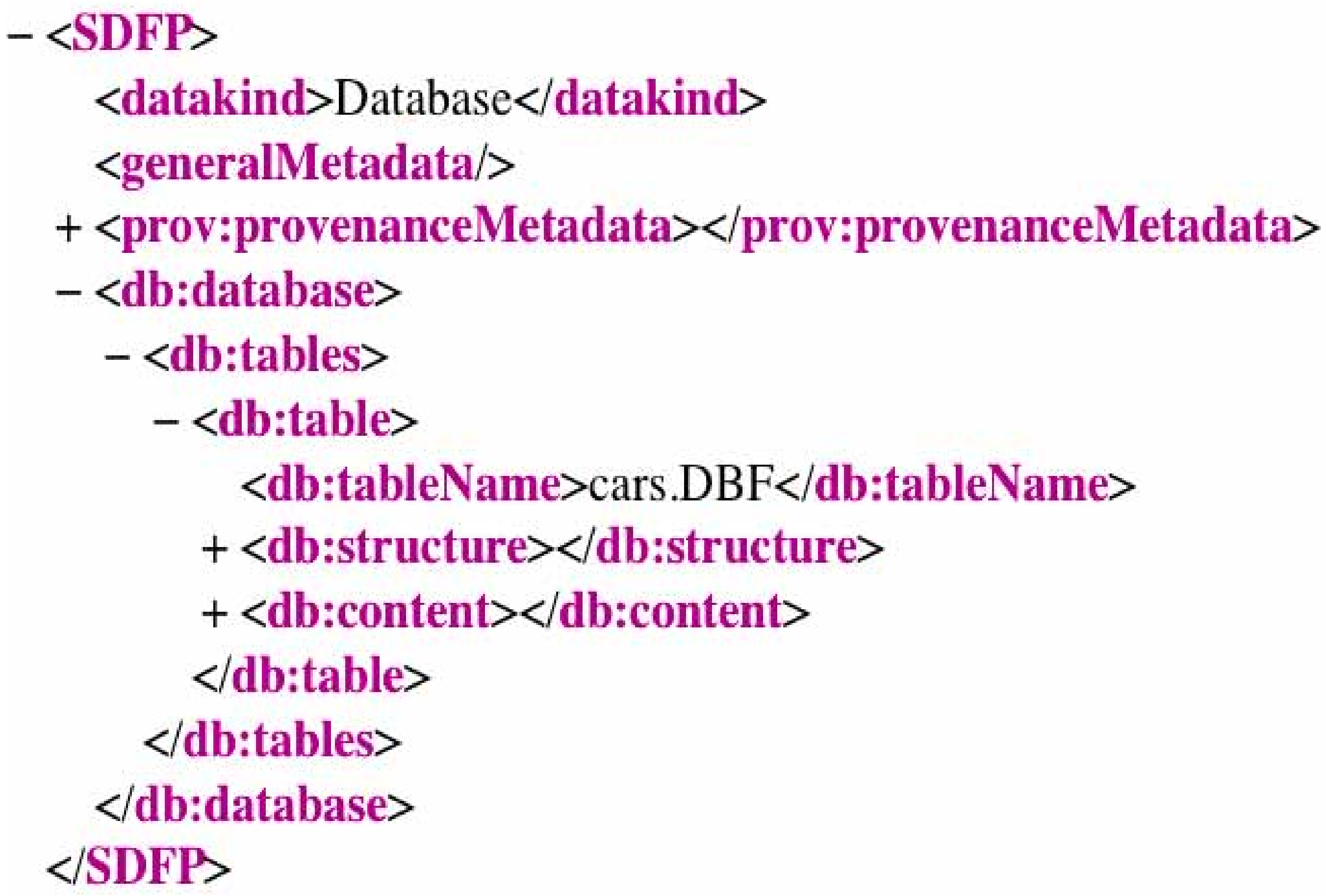
3 **Download converted file**

Download

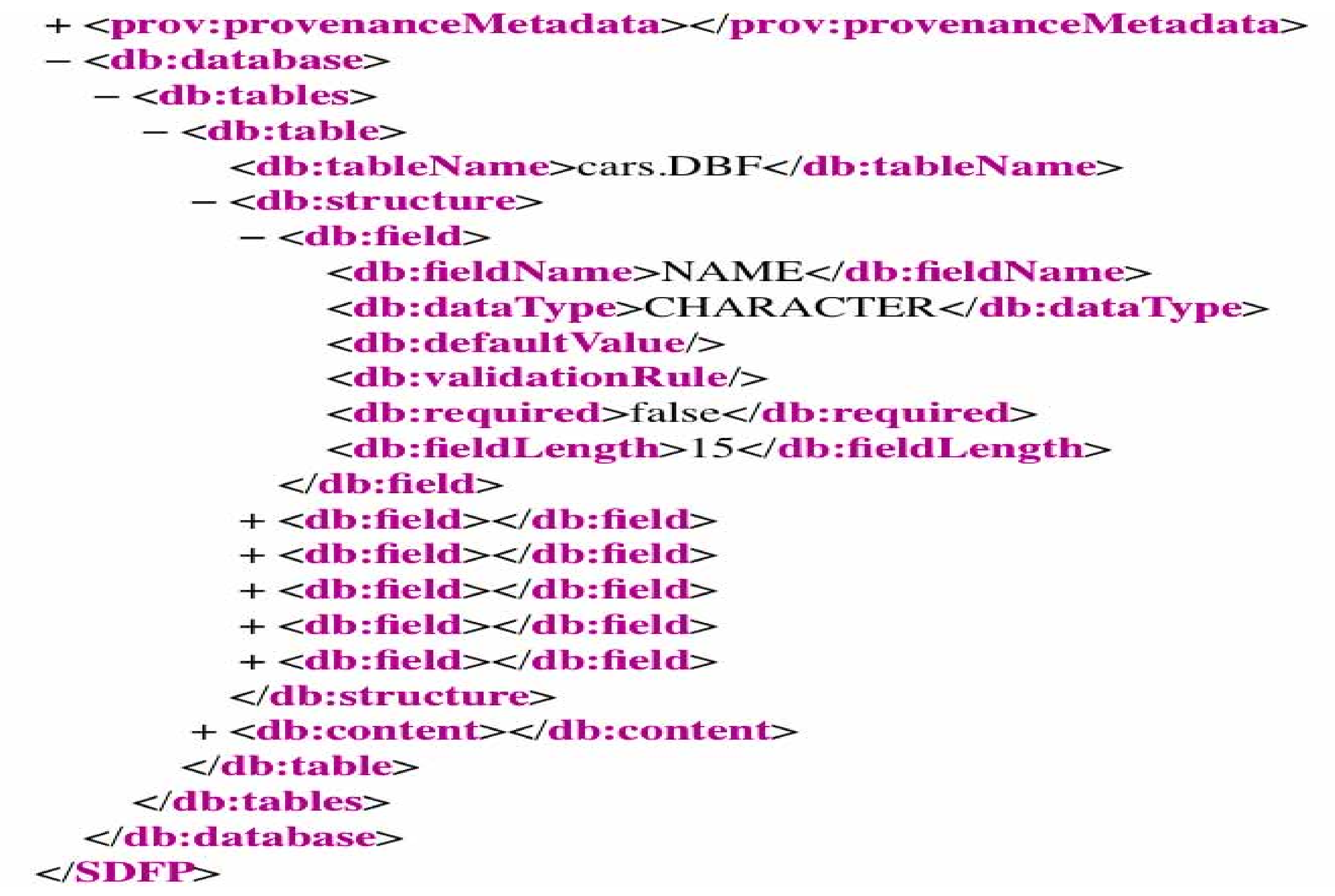
Reported actions of **Batch number: 186**

Job Number	**Report Entry**			
	Source	**Message**	**Date Time**	**Provenance**
187	orchestrator	Starting job 187 for source file file:/tmp/mixed-file-utils-5425351014437147072.temp	Sun 27/09/2009 11:30:10 Show Plugin Status	
	orchestrator	Source file:/tmp/mixed-file-utils-5425351014437147072.temp has file type application/binary;type=dbf	Sun 27/09/2009 11:30:10 Show Plugin Status	
	orchestrator	Fetched 3 plugins	Sun 27/09/2009 11:30:10 Show Plugin Status	
	orchestrator	Using plugin nl.knaw.dans.mixed.converters.convert-dbf-to-sdfp for conversion	Sun 27/09/2009 11:30:10 Show Plugin Status	
	orchestrator	Successfully converted to file:/home/janm/Temp/mixed-job-187-.xml	Sun 27/09/2009 11:30:12 Show Plugin Status	

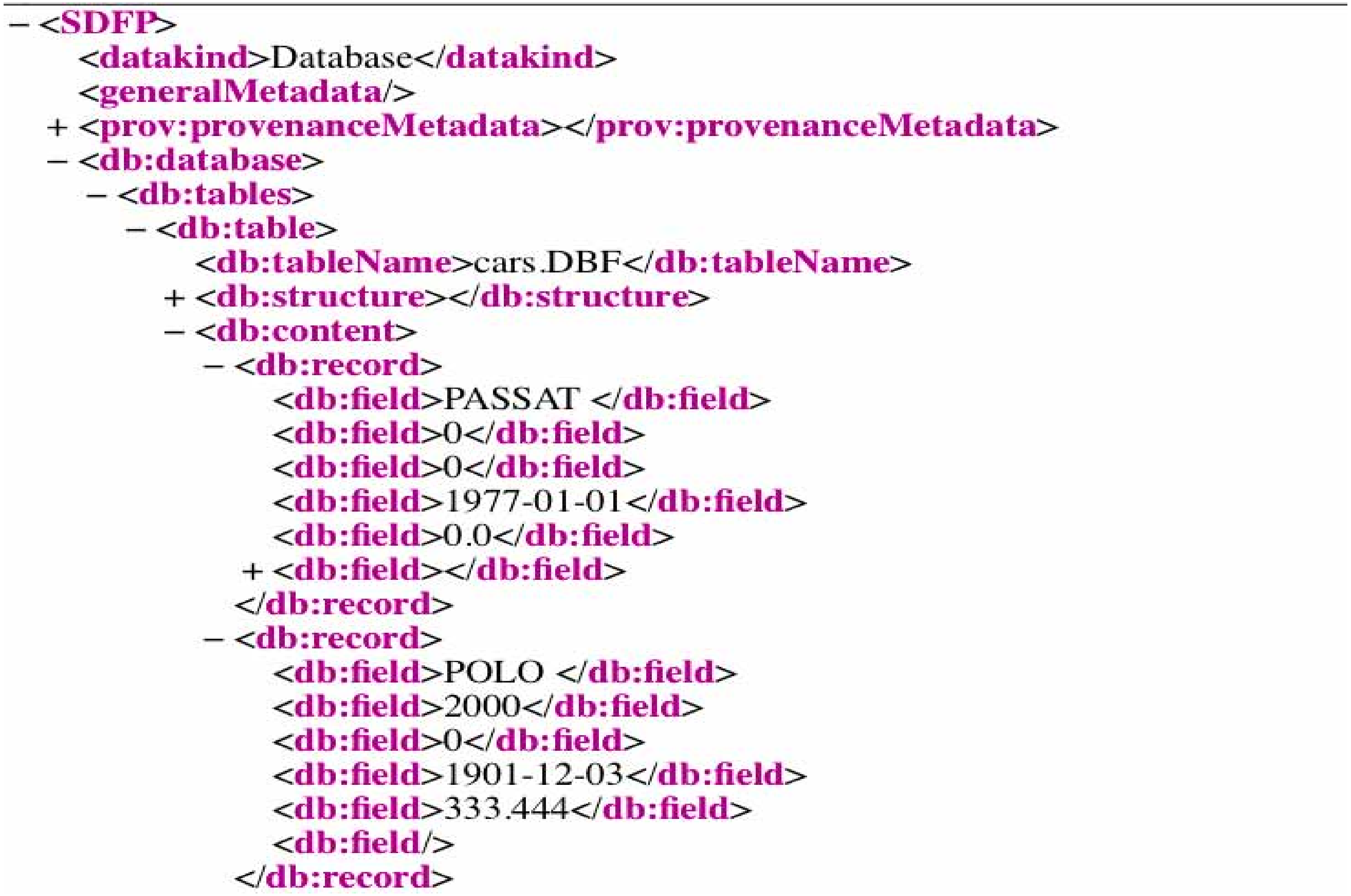
MIXED Version: 1.0-beta



```
– <SDFP>
    <datakind>Database</datakind>
    <generalMetadata/>
  + <prov:provenanceMetadata></prov:provenanceMetadata>
  – <db:database>
    – <db:tables>
      – <db:table>
          <db:tableName>cars.DBF</db:tableName>
        + <db:structure></db:structure>
        + <db:content></db:content>
        </db:table>
      </db:tables>
    </db:database>
  </SDFP>
```



```
+ <prov:provenanceMetadata></prov:provenanceMetadata>
- <db:database>
  - <db:tables>
    - <db:table>
        <db:tableName>cars.DBF</db:tableName>
      - <db:structure>
        - <db:field>
            <db:fieldName>NAME</db:fieldName>
            <db:dataType>CHARACTER</db:dataType>
            <db:defaultValue/>
            <db:validationRule/>
            <db:required>false</db:required>
            <db:fieldLength>15</db:fieldLength>
          </db:field>
        + <db:field></db:field>
        + <db:field></db:field>
        + <db:field></db:field>
        + <db:field></db:field>
        + <db:field></db:field>
        </db:structure>
      + <db:content></db:content>
      </db:table>
    </db:tables>
  </db:database>
</SDFP>
```



```
– <SDFP>
    <datakind>Database</datakind>
    <generalMetadata/>
  + <prov:provenanceMetadata></prov:provenanceMetadata>
  – <db:database>
    – <db:tables>
      – <db:table>
          <db:tableName>cars.DBF</db:tableName>
        + <db:structure></db:structure>
        – <db:content>
          – <db:record>
              <db:field>PASSAT </db:field>
              <db:field>0</db:field>
              <db:field>0</db:field>
              <db:field>1977-01-01</db:field>
              <db:field>0.0</db:field>
            + <db:field></db:field>
            </db:record>
          – <db:record>
              <db:field>POLO </db:field>
              <db:field>2000</db:field>
              <db:field>0</db:field>
              <db:field>1901-12-03</db:field>
              <db:field>333.444</db:field>
              <db:field/>
            </db:record>
```

“Smart migration”

= Conversion upon ingest of specific kinds of data formats (such as spreadsheets and databases) to an intermediate generic format expressed in the XML data format. Upon dissemination the file is converted from this generic format into a current format of choice.

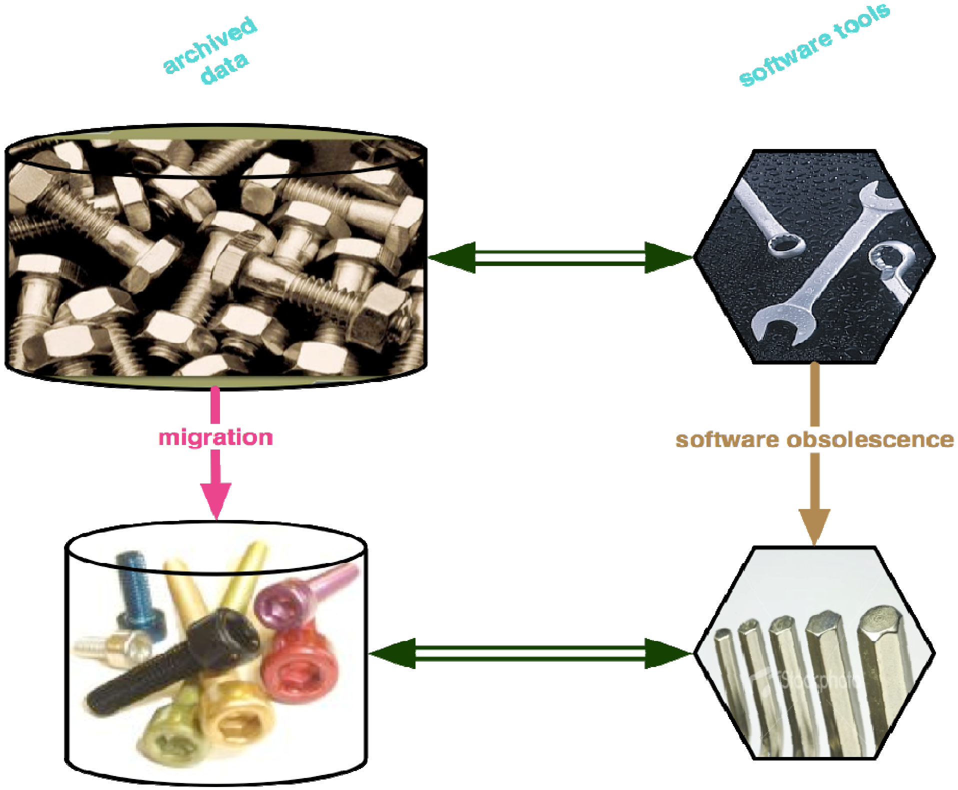
Features of “smart migration”

- It accepts the usable formats of today
- It converts upon ingest into long-term preservation formats
- It converts upon dissemination into usable formats of the future
- No legacy formats required
- No need for a succession of migration between vendor formats over time
- There is always a reusable / data minable copy of the data in the archive

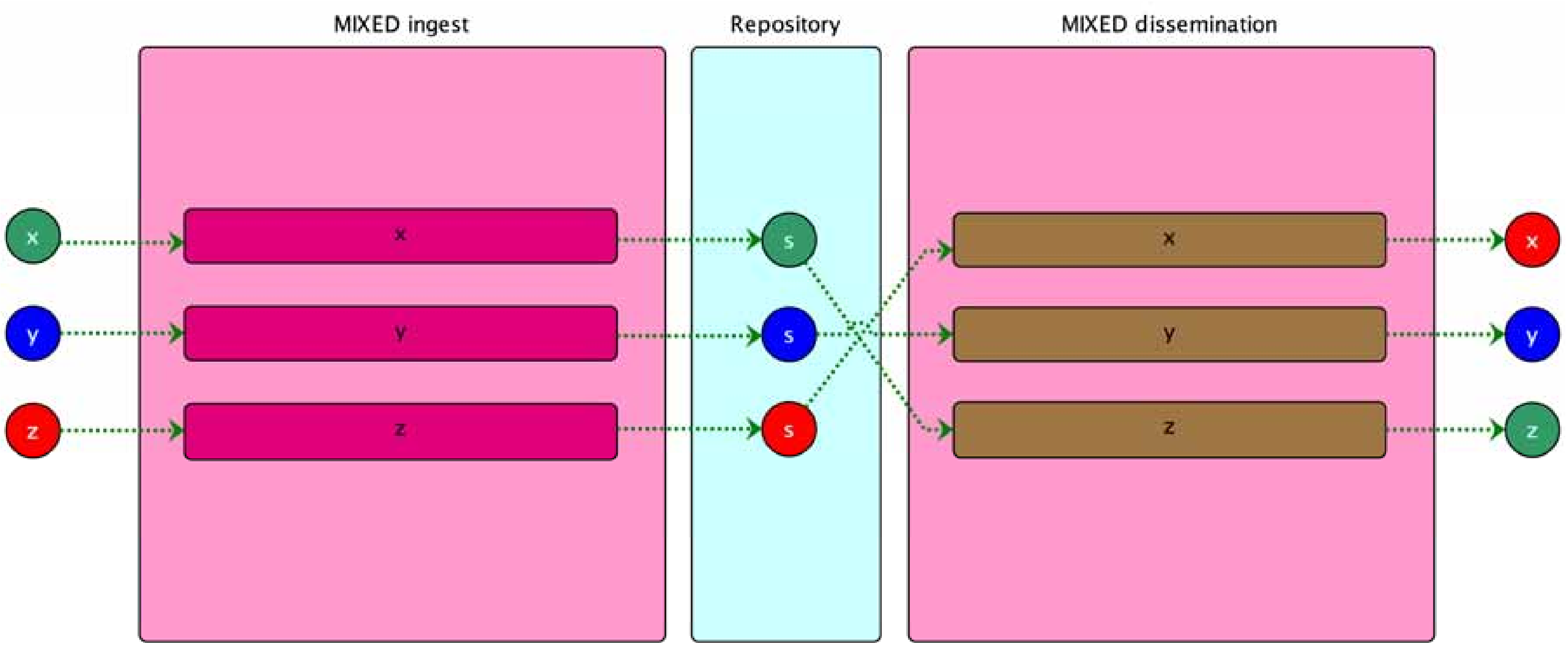
Data Archiving and Networked Services
DANS

MIXED
MIGRATION TO INTERMEDIATE XML FOR ELECTRONIC DATA

bits and tools



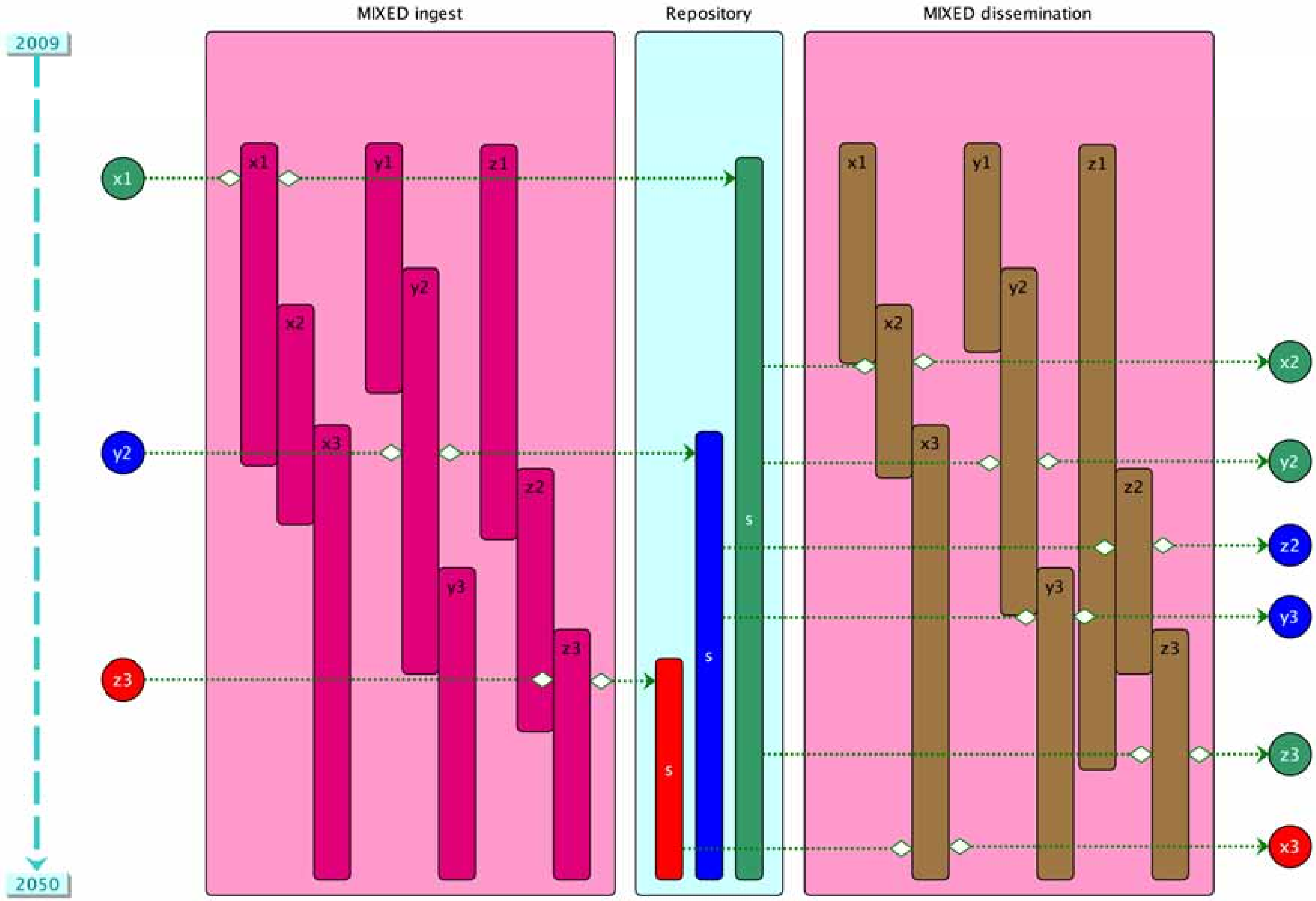
Smart migration - snapshot



Data Archiving and Networked Services
DANS

MIXED
MIGRATION TO INTERMEDIATE XML FOR ELECTRONIC DATA

timeline



SDFP

“Standard Data Formats for Preservation”

- Defines the features of the intermediate XML data format. It is an umbrella format
- Contains sets of XML schemas for various significant data kinds and builds on existing XML representations of file formats
- MIXED: concentrates on tabular data (spreadsheets and databases)
- New data kinds will be added

Current scope of MIXED

Content (semantics)

databases

data model

data itself

spreadsheets

cell positions

values

formulas

Aspects that didn’t make it

presentation details

fonts

forms

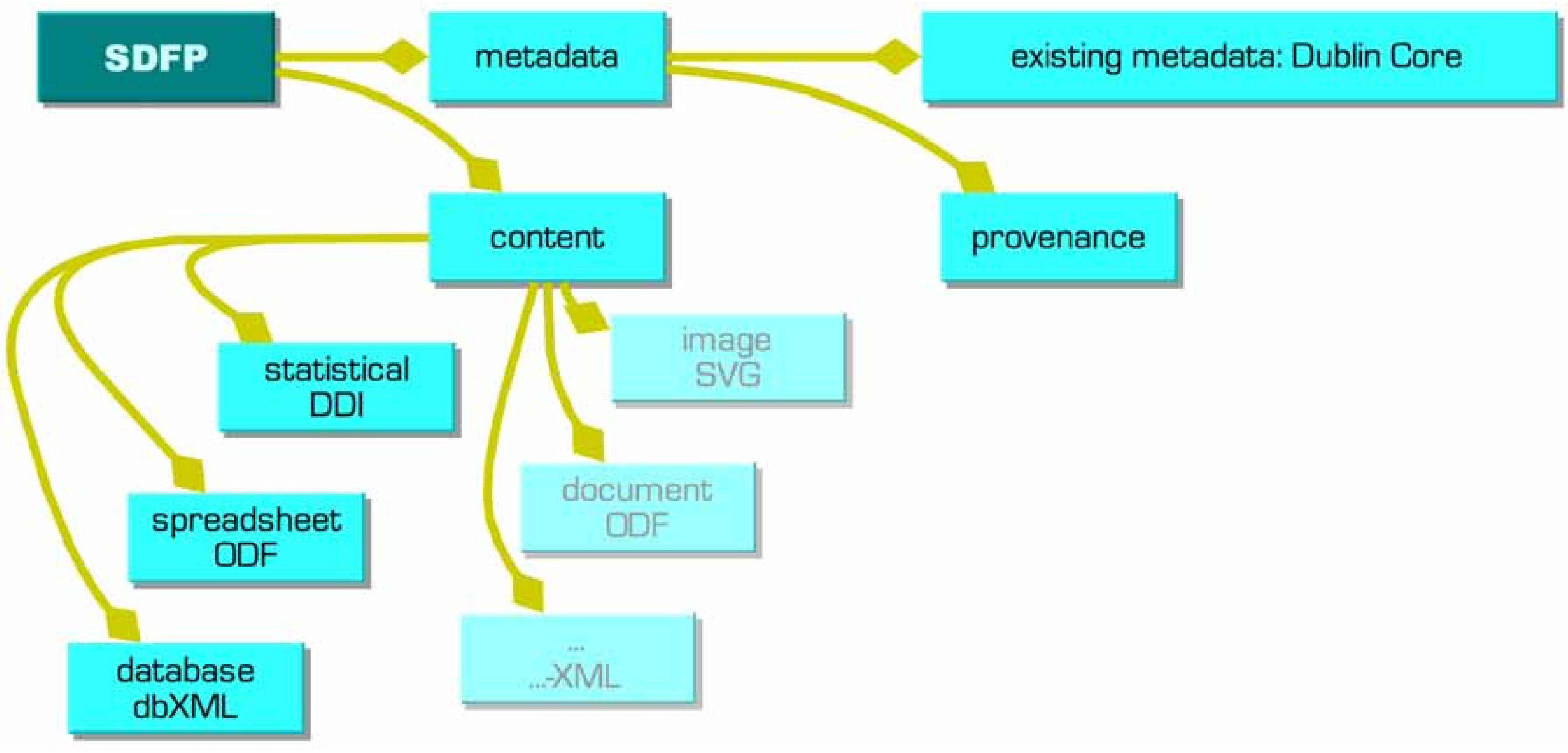
action details

update, insert, delete

stored procedures

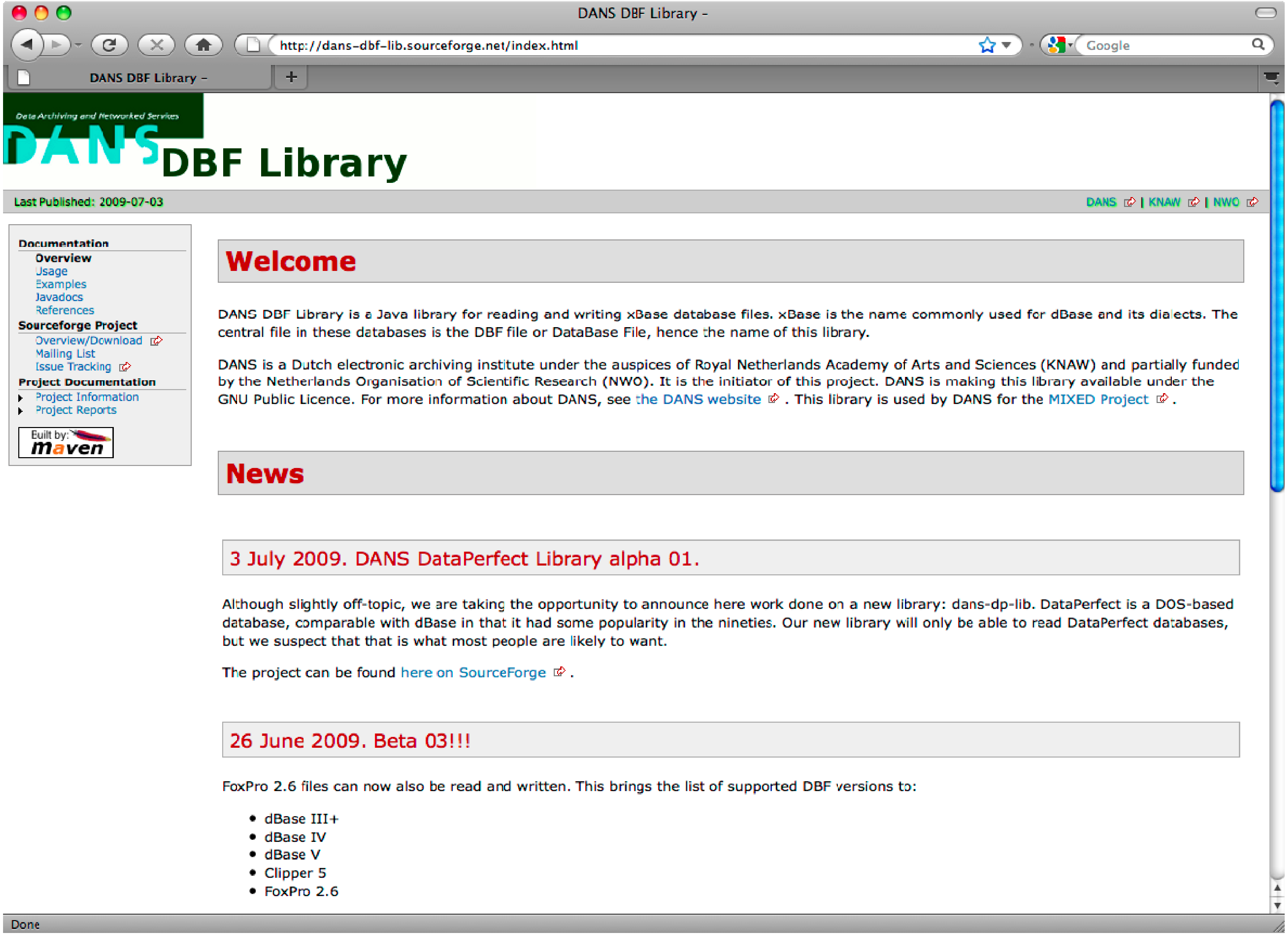
triggers

SDFP as umbrella



MIXED software

- Generic framework with conversion plug-ins
- Open Source
- Several interfaces possible: web console, command line tool, web service, …
- Conversion plug-ins easy extensible / updateable
- Building block in preservation workflow



DBF Library

Last Published: 2009-07-03

Welcome

DANS DBF Library is a Java library for reading and writing xBase database files. xBase is the name commonly used for dBase and its dialects. The central file in these databases is the DBF file or DataBase File, hence the name of this library.

DANS is a Dutch electronic archiving institute under the auspices of Royal Netherlands Academy of Arts and Sciences (KNAW) and partially funded by the Netherlands Organisation of Scientific Research (NWO). It is the initiator of this project. DANS is making this library available under the GNU Public Licence. For more information about DANS, see the DANS website . This library is used by DANS for the MIXED Project .

News

3 July 2009. DANS DataPerfect Library alpha 01.

Although slightly off-topic, we are taking the opportunity to announce here work done on a new library: dans-dp-lib. DataPerfect is a DOS-based database, comparable with dBase in that it had some popularity in the nineties. Our new library will only be able to read DataPerfect databases, but we suspect that that is what most people are likely to want.

The project can be found here on SourceForge .

26 June 2009. Beta 03!!!

FoxPro 2.6 files can now also be read and written. This brings the list of supported DBF versions to:

- dBase III+
- dBase IV
- dBase V
- Clipper 5
- FoxPro 2.6

MIXED

MIGRATION TO INTERMEDIATE XML FOR ELECTRONIC DATA

```
ERROR: stackunderflow
OFFENDING COMMAND: ~

STACK:
```