Document Liberation
Own your content

Document Liberation Project

One Year After

Валёк Филиппов, Reverse Engineer
Fridrich Štrba, Straight Engineer

Document
Liberation
Own your content

The Project

History

- Launched officially on LGM 2014
- Group working on file-formats within LibreOffice since the beginning
    - GSoC 2011 – import filter for Visio file-formats (libvisio)
    - During the year of 2012 – import filter for CorelDraw file-formats (libcdr)
    - GSoC 2012 – import filter for Microsoft Publisher (libmspub)
    - GSoC 2014 – import filter for Adobe PageMaker (libpagemaker)
    - Gates of Heaven opened in 2013 and 2014
    - More to come...

Beyond LibreOffice

- Clear feeling that this is bigger then LibreOffice itself
  - Feedback from conferences
  - Approached by other projects with a lot of interest
- Reuse by other projects
  - Inkscape
  - Calligra
  - Scribus
- A service of the LibreOffice community to the wider FOSS world
  - We receive
  - We give back

Philosophy

- Ownership of documents
  - We believe that documents and their content belong to their creators, not software vendors
- Access to documents
  - We believe that access to content you own should not be hindered by the fact that the application that created it is not maintained any more or that the application does not work on the particular operating system that you use
- Role of open standards
  - We believe that use of truly open and free standards for encoding digital content is the only long-term guarantee that a user's digital content will never be beholden to a single vendor
- Transitory period
  - We believe that implementation of Free and Open Source Software that can read proprietary file-formats is the best solution to escape vendor lock during the transition period to truly open and free standards

Our Mission

- File-format understanding
  - Our mission is to try to understand the structure and details of proprietary, undocumented file-formats
- FOSS parser implementations
  - Our mission is to use the understanding of the file-formats to implement FOSS libraries that are able to parse such documents and extract as much information as possible from them
- ODF eco-system
  - Our mission is to use our existing framework to encode this data in a truly free and open standard file-format: the Open Document Format

Software Framework

- Librevenge
    - APIs and general-use types
- Libodfgen
    - Generators of Open Document files from librevenge APIs
- Parser libraries
    - Libwpd, libwpg, libvisio, libcdr, libmspub, libetonyek, libfreehand, libe-book, libmwaw, libpagemaker,...
    - Parsing file-format && information processing
- Writerperfect
    - Command-line tools to convert to ODF

librevenge-stream

- RVNGInputStream interface
    - Virtual interface allowing stream abstraction
- Several implementations:
    - RVNGFileStream
        - Implementation using file name
    - RVNGStringStream
        - Implementation using a buffer of data
    - RVNGDirectoryStream
        - Accesses a directory structure as if it was a structured document
- OLE2 and ZIP documents handled transparently
    - No need to know what is the container type
    - Gives the responsibility to the implementers!

librevenge-generators

- Useful implementations of the different interfaces
- Raw Generators
  - Implementations of the different librevenge interfaces
    - printing callbacks called and properties passed
  - Used for regression testing
- CSV generator for spreadsheets, HTML, Text generators
- SVG generators
  - Exception: SVG generator for drawings
    - Included in librevenge core library
    - Historical reasons

libodfgen

- Generators for OpenDocument from librevenge interfaces
- **OdtGenerator** class
  - Implementations of *RVNGDocumentInterface*
- **OdgGenerator** class
  - Implementation of *RVNGDrawingInterface*
- **OdpGenerator** class
  - Implementation of *RVNGPresentationInterface*
- **OdsGenerator** class
  - Implementation of *RVNGSpreadsheetInterface*
- **OdfDocumentHandler** interface
  - SAX-like interface to output XML in a generic way

writerperfect

- Command-line tools linking the components together
  - RVNGInputStream implementation
    - Librevenge-stream
  - Different generators
    - Libodfgen – ODF generator
    - Libepubgen – EPUB generator
    - Librvngabw – ABW generator
  - Different parser libraries
    - libvisio, libcdr, libmspub, libetonyek, libwpd, libwpg,....
- Generates Open Document files in two ODF flavours
  - Flat ODF
  - Package (zipped) ODF

Advantage of the design

- Parser libraries independent and self-contained
  - Much easier life of filter writers
    - Enough to focus on the structure of document to parse
    - Call the interface callbacks that one needs
  - Avoid sucking in unrelated libraries
    - Librevenge itself and libodfgen have only boost as build-time dependency
    - No need to link text-related libraries in drawing application
- Considerable reduction of code duplication
  - Less risk to have bugs fixed in one place and hanging around in another
  - Faster to start a library skeleton

Document Liberation

Own your content

Year In Review

Librevenge release

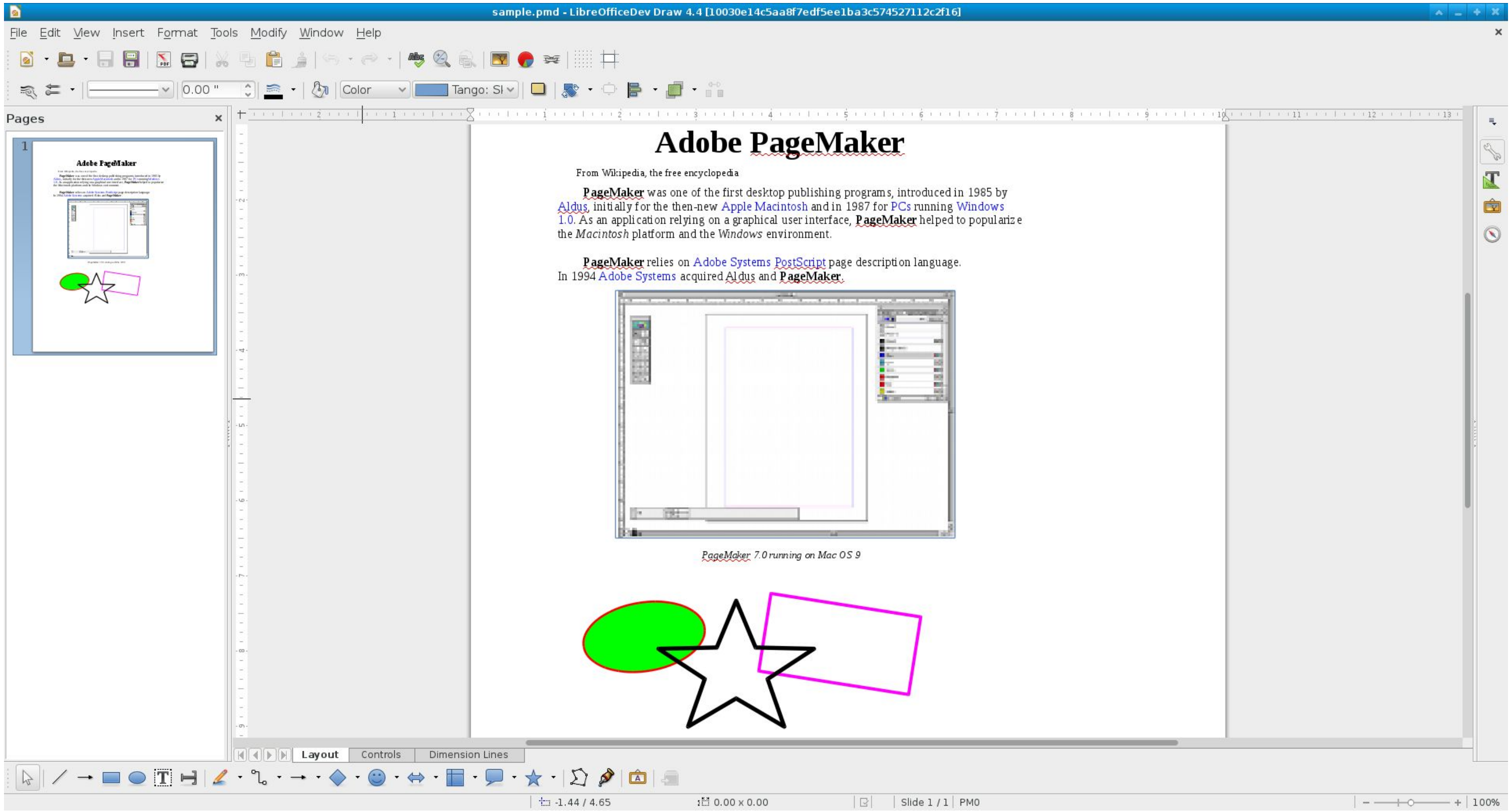
- May 2014
- All libraries were ported to it
- Provided patches to upstream project using our libraries
  - Inkscape
  - Calligra
  - Abiword
  - LibreOffice

New generator libraries

- **libepubgen**
  - Generate EPUB files from text files
  - Generate EPUB files from presentations
  - Generate EPUB files from drawings.
- **Librvngabw**
  - **Try to pronounce it!**
  - Generate Abiword files from different text files

Aldus/Adobe PageMaker

- GSoC 2014 Project, by Anurag Kanungo
- Version 6 and 7



sample.pmd - LibreOfficeDev Draw 4.4 [10030e14c5aa8f7edf5ee1ba3c574527112c2f16]
File Edit View Insert Format Tools Modify Window Help
0.00 "
Color
Tango: Sl
Pages
Layout
Controls
Dimension Lines
-1.44 / 4.65
0.00 x 0.00
Slide 1 / 1
PM0
100%

Adobe PageMaker

From Wikipedia, the free encyclopedia

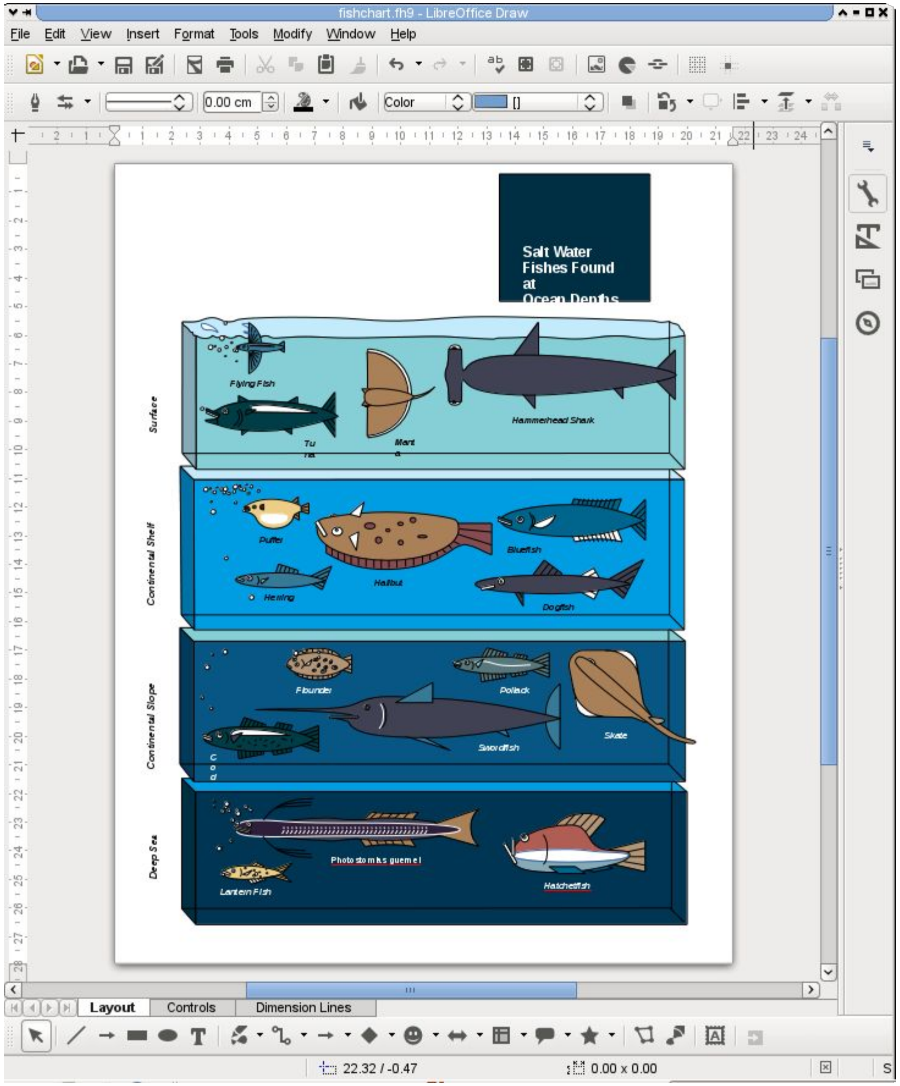
PageMaker was one of the first desktop publishing programs, introduced in 1985 by Aldus, initially for the then-new Apple Macintosh and in 1987 for PCs running Windows 1.0. As an application relying on a graphical user interface, **PageMaker** helped to popularize the *Macintosh* platform and the *Windows* environment.

PageMaker relies on Adobe Systems PostScript page description language.
In 1994 Adobe Systems acquired Aldus and **PageMaker**.

PageMaker 7.0 running on Mac OS 9

Freehand

- One of the most messy file-formats to support
  - We understood more about the file structure
  - Need to be able to parse each record in order to go to the next one.
- Covering currently versions 3 to 11
  - Added simple fills
  - Added text support



Document Liberation
Own your content

UNESCO PERSIST Project

- Under the aegis of the *Memory of the World Prograr*
  - High level policy dialogue on digi preservation between
    - heritage institutions,
    - governments and
    - ICT industry
- Consultative meeting of experts on the PERSIST Project
  - April 20-21, 2015 in Paris
  - Invited as experts in digital preservation
  - Became part of the technical task force

Document Liberation
Own your content

YOU can be part of this!

Ways to contribute

- Code development
  - Contribute to one of our existing libraries, or
  - Start a new one
- Understanding and documenting file-formats
  - OLEToy
    - Preferred way to visualize documents
    - Need a bit of knowledge of Python
- Preparation of sample documents
  - Need to access a generating application
  - Important for regression testing

Future file-formats to import?

- Google Summer of Code 2015
  - We will mentor this time Apple Numbers import library
- Several formats can be improved
  - Freehand
  - Adobe PageMaker
  - MS Publisher 2.0
- Several formats ready for straight engineering
  - Apple Pages
  - Zoner Draw

Document
Liberation
Own your content

Thank you!

www.documentliberation.org
@DocLiberation