Wikipedia Offline

a technical framework

Overview

- Offline Usage
- Technical Requirements
- Different Approaches
- ZenoReader
  - ZenoReader goes Open Source
- Further Development
- ZenoIndexer
- TODO

About Me

- Manuel Schneider from Germany
- in Wikipedia since 2004
- Member of Wikimedia Deutschland and Wikimedia CH
- not really an editor but supporting good ideas and projects – eg. DVD online distribution
- contact person between DirectMedia publisher and tntnet developer

Offline Usage

- mobile users
  – on PDA, Smartphones...
- education in the 3rd world
  – on OLPC, Linux4Africa...
- other users without internet access
- commercial products
  – like german Wikipedia DVD, Kiwix

Technical Requirements

- platform independant
  – should run on Linux, Windows, mobile hardware
- small footprint
  – should run with minimal CPU and RAM usage
- highly compressed data format
  – to store Gigs of text and images
- quick search engine

Different Approaches

- Lampix Wikipedia CD
  – using SQL dumps and MediaWiki, Apache and MySQL server on a Live CD

- Digibib Wikipedia DVD
  – using proprietary Digibib format and reader of DirectMedia, complex conversion

- Kiwix Wikipedia CD
  – selected articles from HTML dump, good search engine, own GUI

Different Approaches

- Wikipedia DVD with ZenoReader
  - creation of a new file format
  - compressed HTML dump
  - new reader software as webserver
  - developed and supported by DirectMedia

ZenoReader

- file format description:
  - header
    - file format version, pointer to index
  - index
    - list of all contents with pointer and length
  - contents
    - content, mime-type, compression-type
- for details see http://wiki.directmedia.de/ZenoReader/Library

ZenoReader

- a completely new approach
- a flexible and yet very straight approach
- DirectMedia implementation (2006/2007):
  - still proprietary as no open source reader was available
  - still platform dependent (Windows)

ZenoReader goes Open Source

- an open source reader application must be implemented:
- tntnet offers the perfect framework for webapplications:
  - C++ webserver library
  - extremely small footprint
  - multi-threaded and extremely fast
  - powerful programming language
  - generic libraries are usable

ZenoReader goes Open Source

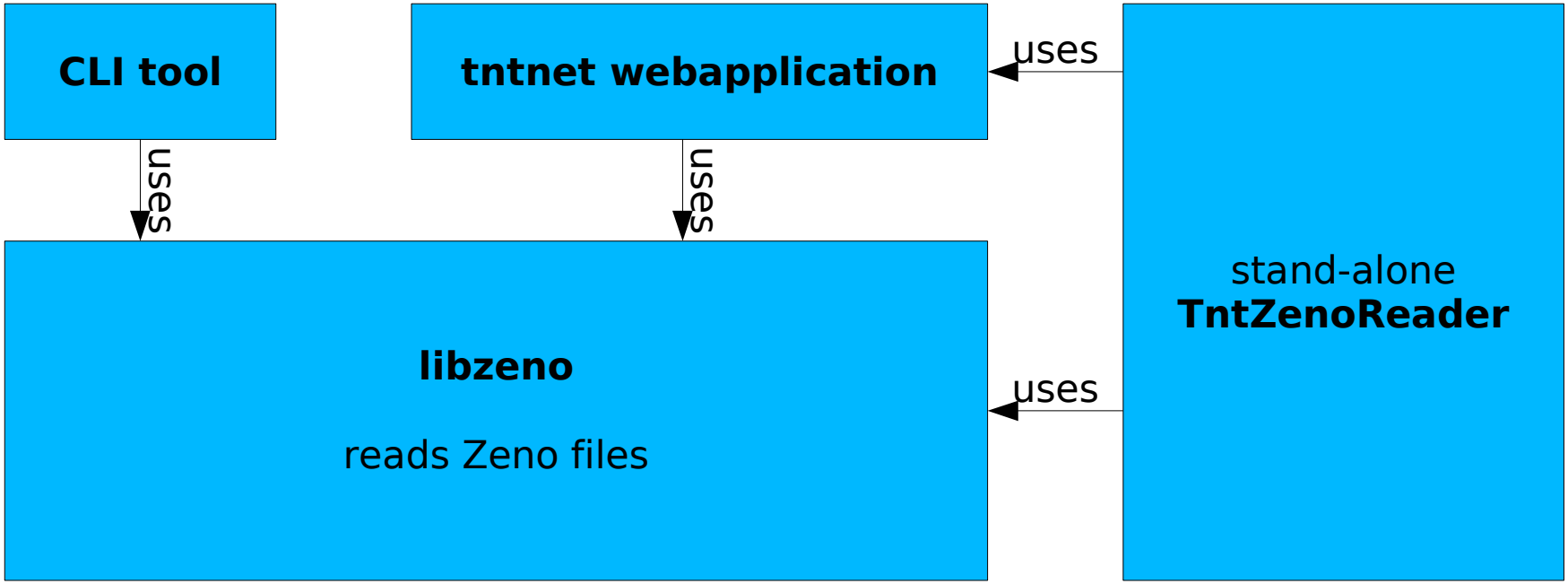
- tntnet developer convinced that a free ZenoReader would be a good demo application
- first prototype was ready after just a few weeks – the first approach of a free ZenoReader (among others) which really survived
- DirectMedia convinced to include a TntZenoReader in the next edition of the DVD
- DirectMedia sponsors the developer – full donation to Wikimedia Deutschland

ZenoReader goes Open Source

- in May 2007 a usable TntZenoReader was available in tntnet svn
- new ideas from the tntnet developer influenced the next edition
- DirectMedia ported the TntZenoReader to Mac OS X
- in November 2007 the lastest edition 2007/2008 was released – including TntZenoReader with binaries for Linux, Mac OS X and source code

ZenoReader goes Open Source

- software architecture:



Further Development

- a prototype of Kiwix already uses libzeno
- Moulin Wiki could be ported to use libzeno
- we're working on packages for:
  - OLPC – swiss OLPC team provides the launcher applett
  - eeePC – Xandros package
  - iPhone – we're signed up for the development kit as soon as it is available for Europe
    – we don't have an iPhone: LOOKING FOR TESTERS!
- all other Unices: If you can compile it yourself, you can use it anytime

Further Development

- to read free content from ZenoFiles, we need ZenoFiles
- work on an open source ZenoIndexer is done
- DirectMedia permits us to reimplement the software as free software (GPL)
- some improvements will be included, but without changing the file format

ZenoIndexer

- idea:

1. create a list of articles to be downloaded
2. ZenoIndexer downloads the articles from the Wiki and cleans them (extracts content portion, removes boxes, fixes links)
3. the usable content is put into a database
4. after all content is fetched, the contents can be prepared if necessary
5. the prepared content is being compressed and dumped into a new ZenoFile

ZenoIndexer

- technical improvements:
  - general usage of unicode (utf-8)
  - read-ahead to improve access time on DVDs / CDs
  - switching compression to LZMA
    – save another 50% compared to gzip
  - ZenoIndexer uses tntdb which is the frontend to several databases (MySQL, SQLite, Postgres, ORACLE)

TODO

- Technical work to be done:
  - test TntZenoReader for bugs
  - compile ready-to-use packages for different Unices / Linux distributions
  - help with porting to the iPhone or other embedded platform
  - help developing the ZenoIndexer
- all software is available in the tntnet svn!

  http://www.tntnet.org/tntreader.html

TODO

- Legal work:
- reliability: Who's getting sued for unsuitable data in the distribution?
- find a publisher (company) rather than do it by yourself
- keep it open – the software, not only the contents

Discussion

- Ideas
- Questions
- Contributions