



# What You Always Wanted To Know About Tesseract

0. Tutorial Contents

*Ray Smith, Google Inc.*

# Tutorial Contents

1. 8:30-9:00 Introduction/Motivation/History
2. 9:00-9:20 Architecture and Data Structures
3. 9:20-9:45 Training Tesseract
4. 9:45-10:00 Downloading/Building + Coffee Break
5. 10:00-10:50 Features and Character Classifier
6. 10:50-11:30 Character Segmentation, Language Models and Beam Search
7. 11:30-12:00 Page Layout Analysis
8. 12:00-12:30 Modernization Efforts

# Downloading the Slides

The Tesseract Home Page is at: <https://code.google.com/p/tesseract-ocr/>



An OCR Engine that was developed at HP Labs between 1985 and 1995... and now at Google.

Project Home Downloads Wiki Issues Source

Summary People

Project information

Started by 3631 users  
Project leads

Code license  
Apache License 2.0

Labs  
OCR, Utility, CPaaSPlus, Google

Members  
theragati@gmail.com  
devile\_@gmail.com  
tessera\_@gmail.com  
12 contributors

Featured

Downloads  
tesseract-3.02-wc32-ib-includes.dll.zip  
tesseract-ocr-3.02-wc2008.zip  
tesseract-ocr-3.02-wc32-portable.zip  
tesseract-ocr-3.02-08w-rtm1.jar.zip  
tesseract-ocr-3.02-02-16r.zip  
tesseract-ocr-API-Example-v3008.zip  
tesseract-ocr-portable-3.02-02.exe  
Show all 3

Wiki pages  
JpFcity  
Addons  
FAQ  
Installation  
ReadMe  
TrainingTesseract3  
Show all 3

Links

External links  
Tesseract Forum  
OCRequis  
Design Documents  
Tesseract Developers

Groups  
Tesseract Forum  
Issues List  
Bugs/Feature Requests  
Developers Forum

Tesseract is probably the most accurate open source OCR engine available. Combined with the [Leptonica Image Processing Library](#) it can read a wide variety of image formats and convert them to text in over 60 languages. It was one of the top 3 engines in the 1995 UNLV Accuracy test. Between 1985 and 2008 it had little work done on it, but since then it has been improved extensively by Google. It is released under the [Apache License 2.0](#).

- [ReadMe](#) - Installation and usage information.
- [Compiling](#) - How to build Tesseract on a variety of platforms.
- [FAQ](#) - Common questions and problems. Please check before [filing a bug](#) or [consulting the forum](#).
- [Too many errors?](#) - See the guidance on getting the best out of Tesseract.

## Supported Platforms

Tesseract works on Linux, Windows (with VC++ Express or CygWin) and Mac OS/X. See the [Downloads](#) for more details and install instructions. It can also be compiled for other platforms, including Android and the iPhone, though these are not as well tested platforms. See also the [Wiki](#) page for other projects using Tesseract on various platforms.

If you're interested in supporting other platforms or languages, please get in touch with Ray Smith or the [Developers](#).

## A Note about Downloads

With the discontinuation of downloads at code.google.com, new source downloads will be moved to [GoogleCloud](#). Other download folders will be setup as new files are uploaded, and the original Downloads page will go away. During the transition, other downloads will still be found at the [Old Downloads](#) page.

## Roadmap

Version 3.03 release candidate is now available (source only so far) for download and contains many new features. (See the [ReleaseNotes](#) for a full list.) Please check out the [Roadmap](#) before going to [Downloads](#) as you need more than one file. Even the windows executables listed are incomplete as [language files](#) are required. Most notable new features:

- PDF output.
- New Renderer for extracting detailed recognition information at a document level.

Version 3.03 ships with recent Linux distributions such as Ubuntu 14.04.

Version 3.02 ships with Ubuntu 12.04

## Core Developers

The core developer on the project is Ray Smith (theragati@).

In related work, Thomas Breuel (tbreuel) and Ilya Mochelov (mochelov) work on the [OCRequis](#) project, which also provides layout analysis and statistical language modeling.

Most of the work on Tesseract is sponsored by Google.

Click on the  
GoogleDrive  
link

Download TutorialSlides.tar.  
gz