Home

`blocked URL`

General Information

- Tika Website
- Download latest Tika Release
- Tika mailing lists: Sign-up
- Tika Support
- TikaResources - Articles, books, podcasts, etc. on using Tika
- Troubleshooting Tika
- 3rd party parser plugins
- DeveloperResources
- Using Tika Server - How to deploy Tika as a RESTful Service.
- API Bindings for Tika
- Logging

Migrating to Tika 2.0.0

- Migrating to 2.0.0

Contributing to the wiki

To help avoid spam, in common with many other ASF wikis, the Tika wiki is only editable by known accounts.

If you would like to help out with the Tika wiki, add a new page, or work on an existing one, please first create a wiki account. With that done, drop an email to the user list or the dev list with your wiki username asking for access, and generally within a few hours you'll be able to edit away from then on!

Committer Info

- UsingGit - Information on Tika's configuration management using Git.
- Release Process - Info on releasing Tika
- ThirdPartySonaType - A guide to staging and deploying third party jars on Sonatype OSSRH (OSS Repository Hosting) for subsequent use within Tika parser wrappers
- VirtualMachine - a virtual machine hosted by Rackspace that allows an instance of Tika Server to run for public testing. Set up by Tim Allison et al.

User Notes

- Using Tika Server - How to deploy Tika as a RESTful Service.
- ModifyingContentWithHandlersAndMetadataFilters How to configure limits and modify parsed content with the AutoDetectParserConfig, custom ContentHandlers, metadata filters and metadata write filters.
- API Bindings for Tika - Using Tika from additional languages and frameworks.
- PostingManyFilesToExtractingRequestHandler - How to post many files to the Extracting Request Handler (Tika) in Solr.
- IntegratingTikaWithExtractingRequestHandler - Building the latest Tika and integrating it with the Extracting Request Handler (Tika) in Solr.
- Some stats using Tesseract OCR - some stats from a contributing team (Hyperion Gray) about using TesseractOCR (will be updated with Tika).
- Troubleshooting Tika
- Notes on configuring parsing via the ParseContext
- Notes on Specific Parsers

- Notes on configuring Tika to extract embedded vba and js
- Using the tika-eval Module
- When does Tika need/create a File rather than an InputStream?
- How to Test Your Framework's Handling of Tika Behaving Badly

MIME identification design/implementation

- Bayesian MIME selection - Tika's new Bayesian MIME selector.
- Content-based MIME selection with Byte histograms - Tika's new content/byte histogram MIME detector.

Advanced Content Extraction with Tika - Integration

- Getting Tika up and Running with Pooled Time Series - How to use Tika with the Pooled Time Series video descriptor similarity code.
- Getting Tika up and Running with Apache cTAKES - How to use Tika with Apache cTAKES the clinical text biomedical knowledge extraction framework.
- Getting Tika up and Running with EXIFTool - How to use Tika with EXIFTool.
- Getting Tika up and Running with FFMPEG - How to use Tika with FFMPEG.
- Getting Tika up and Running with the GROBID PDF Journal parser - How to use Tika with the GROBID PDF journal parser.
- Getting Tika up and Running with the GeoTopicParser based on Geonames.org, Lucene, and OpenNLP
- Getting Tika up and Running with OCR - How to use Tika with OCR from Tesseract.
- Getting Tika up and Running with the Geospatial Data Abstraction Library (GDAL) - How to use Tika with GDAL to parse/extract geospatial data files.

Entity Recognition Support

Named Entity Recognition (NER) support

- Getting Tika up and running with Stanford Core NLP and with OpenNLP - How to use Tika with Stanford NER/NLP and with Apache Open NLP.
- Getting Tika up and running with NLTK - How to use Tika with the Python Natural Language Toolkit (NLTK).
- Getting Tika up and running with Grobid Quantities Measurement Parsing - How to use Tika with the Grobid Quantities measurement parser.
- Getting Tika up and running with MIT Lincoln Lab's MIT-nlp Information Extraction (MITIE) toolkit - How to use Tika with MITIE from MIT-NLP.
- Getting Tika up and running with automatic Age Detection from Text - How to use Tika with USC IRDS age detection tools.

Object Recognition (Computer Vision) support

Images

- Getting Tika up and running for Image Visual Recognition - How to use Tika with Tensorflow's Inception-V4 ImageNet for visual recognition of images.
- Getting Tika up and running for Image Visual Recognition using Deep Learning 4J (DL4J) - How to use Tika with Tensorflow's Inception V-3 ImageNet and VGG-16 for visual recognition in pure Java.
- Getting Tika up and running for Computer Vision - Image Captioning - How to use Tika with Tensorflow for combining Computer Vision and NLP to automatically generate captions of images.

Video

- Getting Tika up and running for Video Visual Recognition - How to use Tika with Tensorflow's Inception-V4 ImageNet for visual recognition of videos.

Language Translation

Statistical Machine Translation

- Statistical Machine Translation with Apache Joshua (Incubating) - A guide for leveraging Apache Joshua for language translation via the Tika.translate API.
- Neural Machine Translation powered by Reader Translator Generator toolkit - A guide for RTG integration with Tika.translate API

Design

- MetadataDiscussion - discussions on the design of MIME type detection and parsing for recursive metadata formats (and container formats) like Zip, etc.
- RecursiveMetadata - proposals for dealing with recursive metadata, based on the MetadataDiscussion page.
- Tika JAX-RS Server - documentation on the recently contributed tika-server module.
- Metadata roadmap - Documentation and Discussion about the metadata roadmap for Tika

- Errors and Exceptions - What parsers should output/throw when, for empty/invalid/unsupported files
- Composite Parsers discussion - How to give users sensible+clear control of multiple parsers for a given file type
- Tika 2.0 discussion - Roadmap for changes we would like to make for Tika 2.0
- Tika 2.0 Migration Guide - Guide for migrating to Tika 2.0 (once it is available)

Meetings and Tutorials

- Apache Tika Meetups
- Open Preserve Foundation Talk on Embedded Files

Regression Testing On the Rackspace VM

How to run tika-eval on the VM