**Apache Tika** is a content detection and [analysis](https://en.wikipedia.org/wiki/Content_analysis) framework, written in [Java](https://en.wikipedia.org/wiki/Java_(programming_language)), stewarded at the [Apache Software Foundation](https://en.wikipedia.org/wiki/Apache_Software_Foundation).[[1]](https://en.wikipedia.org/wiki/Apache_Tika#cite_note-1) It detects and extracts metadata and text from over a thousand different [file types](https://en.wikipedia.org/wiki/File_type), and as well as providing a [Java](https://en.wikipedia.org/wiki/Java_(programming_language)) library, has server and command-line editions suitable for use from other programming languages.



**Contents**

* [1History](https://en.wikipedia.org/wiki/Apache_Tika#History)
* [2Features](https://en.wikipedia.org/wiki/Apache_Tika#Features)
* [3Notable uses](https://en.wikipedia.org/wiki/Apache_Tika#Notable_uses)
* [4See also](https://en.wikipedia.org/wiki/Apache_Tika#See_also)
* [5References](https://en.wikipedia.org/wiki/Apache_Tika#References)

History[[edit](https://en.wikipedia.org/w/index.php?title=Apache_Tika&action=edit&section=1)]

The project originated as part of the [Apache Nutch](https://en.wikipedia.org/wiki/Apache_Nutch) codebase, to provide content identification and extraction when [crawling](https://en.wikipedia.org/wiki/Web_crawlers). In 2007, it was separated out, to make it more extensible and usable by [content management systems](https://en.wikipedia.org/wiki/Content_management_systems), other [Web crawlers](https://en.wikipedia.org/wiki/Web_crawlers), and information retrieval systems. The standalone Tika was founded by Jérôme Charron, [Chris Mattmann](https://en.wikipedia.org/wiki/Chris_Mattmann) and Jukka Zitting.[[2]](https://en.wikipedia.org/wiki/Apache_Tika#cite_note-2) In 2011 Chris Mattmann and Jukka Zitting released the Manning book "Tika in Action", and the project released version 1.0.

Features[[edit](https://en.wikipedia.org/w/index.php?title=Apache_Tika&action=edit&section=2)]

Tika provides capabilities for identification of more than 1400 file types from the [Internet Assigned Numbers Authority](https://en.wikipedia.org/wiki/Internet_Assigned_Numbers_Authority) taxonomy of [MIME](https://en.wikipedia.org/wiki/MIME) types. For most of the more common and popular formats,[[3]](https://en.wikipedia.org/wiki/Apache_Tika#cite_note-3) Tika then provides content extraction, metadata extraction and language identification capabilities.

While Tika is written in [Java](https://en.wikipedia.org/wiki/Java_(programming_language)), it is widely used from other languages.[[4]](https://en.wikipedia.org/wiki/Apache_Tika#cite_note-4) The [RESTful](https://en.wikipedia.org/wiki/Representational_state_transfer) server and [CLI Tool](https://en.wikipedia.org/wiki/Command-line_interface) permit non-Java programs to access the Tika functionality.

Notable uses[[edit](https://en.wikipedia.org/w/index.php?title=Apache_Tika&action=edit&section=3)]

Tika is used by financial institutions including the [Fair Isaac Corporation](https://en.wikipedia.org/wiki/Fair_Isaac_Corporation) (FICO),[[5]](https://en.wikipedia.org/wiki/Apache_Tika#cite_note-5) Goldman Sachs,[[6]](https://en.wikipedia.org/wiki/Apache_Tika#cite_note-6) [NASA](https://en.wikipedia.org/wiki/NASA) and academic researchers[[7]](https://en.wikipedia.org/wiki/Apache_Tika#cite_note-7) and by major content management systems including [Drupal](https://en.wikipedia.org/wiki/Drupal),[[8]](https://en.wikipedia.org/wiki/Apache_Tika#cite_note-8) and [Alfresco (software)](https://en.wikipedia.org/wiki/Alfresco_(software))[[9]](https://en.wikipedia.org/wiki/Apache_Tika#cite_note-9) to analyze large amounts of content, and to make it available in common formats using information retrieval techniques.

On April 4, 2016[[10]](https://en.wikipedia.org/wiki/Apache_Tika#cite_note-10) [Forbes](https://en.wikipedia.org/wiki/Forbes) published an article identifying Tika as one of the key technologies used by more than 400 journalists to analyze 11.5 million leaked documents that expose an international scandal involving world leaders storing money in offshore [shell corporations](https://en.wikipedia.org/wiki/Shell_corporation). The leaked documents and the project to analyze them is referred to as the [Panama Papers](https://en.wikipedia.org/wiki/Panama_Papers).