Tika – Content Analysis Toolkit

Apache Tika is a toolkit for detecting and extracting metadata and structured text content from various documents using existing parser libraries.

Apache Tika is an effort undergoing incubation at The Apache Software Foundation (ASF), sponsored by the Apache Lucene PMC. Incubation is required of all newly accepted projects until a further review indicates that the infrastructure, communications, and decision making process have stabilized in a manner consistent with other successful ASF projects. While incubation status is not necessarily a reflection of the completeness or stability of the code, it does indicate that the project has yet to be fully endorsed by the ASF.

See the Apache Tika Incubation Status page for the current incubation status.

Latest News

March 22nd, 2007: Apache Tika project started
The Apache Tika project was formally started when the Tika proposal was accepted by the Apache Incubator PMC.