GrobidJournalParser

The GrobidJournalParser uses the GROBID (or Grobid) GeneRation Of Bibliographic Data machine learning framework to parse PDF documents and to extract structured informations such as title, abstract, authors, affiliations, keywords, etc, from journal publications. The parser has been integrated into Tika. You can follow this guide to get it working on your system.

Installing GROBID

The best approach is to run Grobid via docker.

TLDR: The following command will start the lightweight Grobid image on port 8070 (the default Grobid port):

```
docker run -t --rm --init -p 8070:8070 lfoppiano/grobid:${latest_grobid_version}
```

For more detailed information see https://grobid.readthedocs.io/en/latest/Grobid-docker/.

TIKA configuration

NOTE: We assume the JAR file is downloaded in the `$HOME` directory.

First, we need to create the required configuration files (`GrobidExtractor.properties`) that points to the Grobid service.

You can set it up using the following Github project, and modify the `GrobidExtractor.properties` file accordingly.

1. `cd $HOME && git clone https://github.com/chrismattmann/grobidparser-resources.git`
2. modify the file `grobidparser-resources/org/apache/tika/parser/journal/GrobidExtractor.properties`

Both tika-server and tika-parser-nlp are required for calling Grobid.

Running Grobid with Tika Server

When you start Tika Server, use the following command.

```
java -cp grobidparser-resources:tika-server-standard-2.8.0.jar:tika-parser-nlp-package-2.8.0.jar org.apache.
tika.server.core.TikaServerCli --config grobidparser-resources/tika-config.xml
```

Then, PUT a file to Tika-server as follow (the filename needs to be specified in the headers `-H`):

```
curl -T $PATH_TO_FILENAME -H "Content-Disposition: attachment;filename=$FILENAME.PDF" http://localhost:9998
/rmeta
```

For example:

```
curl -T $PATH/...ICSE06.pdf -H "Content-Disposition: attachment;filename=ICSE06.pdf" http://localhost:9998/rmeta
```

Which will output (if e.g., using `python -m json.tool`):


```
[
    {
        "Author": "End User Computing Services",
        "Company": "ACM",
        "Content-Type": "application/pdf",
        "Creation-Date": "2006-02-15T21:13:58Z",
```

```
        "Last-Modified": "2006-02-15T21:16:01Z",
        "Last-Save-Date": "2006-02-15T21:16:01Z",
        "SourceModified": "D:20060215211344",
        "X-Parsed-By": [
            "org.apache.tika.parser.CompositeParser",
            "org.apache.tika.parser.journal.JournalParser"
        ],
        "X-TIKA:content":
"\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\nProceedings
Template - WORD\n\n\nA Software Architecture-Based Framework for Highly \nDistributed and Data Intensive
Scientific Applications \n\n \nChris A. Mattmann1, 2        Daniel J. Crichton1        Nenad Medvidovic2
Steve Hughes1 \n\n \n1Jet Propulsion Laboratory \n\nCalifornia Institute of Technology \nPasadena, CA 91109,
USA \n\n{dan.crichton,mattmann,steve.hughes}@jpl.nasa.gov \n\n2Computer Science Department \nUniversity of
Southern California  \n\nLos Angeles, CA 90089, USA \n{mattmann,neno}@usc.edu \n\n \nABSTRACT \nModern
scientific research is increasingly conducted by virtual \ncommunities of scientists distributed around the
world. The data \nvolumes created by these communities are extremely large, and \ngrowing rapidly. The
management of the resulting highly \ndistributed, virtual data systems is a complex task, characterized \nby a
number of formidable technical challenges, many of which \nare of a software engineering nature.  In this paper
we describe \nour experience over the past seven years in constructing and \ndeploying OODT, a software
framework that supports large, \ndistributed, virtual scientific communities. We outline the key \nsoftware
engineering challenges that we faced, and addressed, \nalong the way. We argue that a major contributor to the
success of \nOODT was its explicit focus on software architecture. We \ndescribe several large-scale, real-
world deployments of OODT, \nand the manner in which OODT helped us to address the domain-\nspecific challenges
induced by each deployment.  \n\nCategories and Subject Descriptors \nD.2 Software Engineering, D.2.11 Domain
Specific Architectures \n\nKeywords \nOODT, Data Management, Software Architecture. \n\n1. INTRODUCTION ..
snip..",
        "X-TIKA:parse_time_millis": "957",
        "access_permission:assemble_document": "true",
        "access_permission:can_modify": "true",
        "access_permission:can_print": "true",
        "access_permission:can_print_degraded": "true",
        "access_permission:extract_content": "true",
        "access_permission:extract_for_accessibility": "true",
        "access_permission:fill_in_form": "true",
        "access_permission:modify_annotations": "true",
        "created": "Wed Feb 15 13:13:58 PST 2006",
        "creator": "End User Computing Services",
        "date": "2006-02-15T21:16:01Z",
        "dc:creator": "End User Computing Services",
        "dc:format": "application/pdf; version=1.4",
        "dc:title": "Proceedings Template - WORD",
        "dcterms:created": "2006-02-15T21:13:58Z",
        "dcterms:modified": "2006-02-15T21:16:01Z",
        "grobid:header_Address": "Pasadena, CA 91109 USA Los Angeles, CA 90089 USA ",
        "grobid:header_Affiliation": "1 Jet Propulsion Laboratory California Institute of Technology; 2
Computer Science Department University of Southern California",
        "grobid:header_Authors": "Chris A Mattmann 1,2 Daniel J Crichton 1 Nenad  Medvidovic 2 Steve  Hughes 1
",
        "grobid:header_Class": "org.apache.tika.metadata.Metadata",
        "grobid:header_FullAffiliations": "[Affiliation {orgName=Jet Propulsion Laboratory California Institute
of Technology , address=Pasadena, CA 91109 USA},Affiliation {orgName=Computer Science Department University of
Southern California , address=Los Angeles, CA 90089 USA}[Affiliation {orgName=Jet Propulsion Laboratory
California Institute of Technology , address=Pasadena, CA 91109 USA},Affiliation {orgName=Computer Science
Department University of Southern California , address=Los Angeles, CA 90089 USA}]",
        "grobid:header_Keyword": "\"D2 Software Engineering, D211 Domain Specific Architectures\"",
        "grobid:header_TEIJSONSource": "{\"TEI\":{\"text\":{\"xml:lang\":\"en\"},\"teiHeader\": ..snip",
        "grobid:header_TEIXMLSource": "<?xml version=\"1.0\" encoding=\"UTF-8\"?>\n<?xml-model href=\"
file:///Users/mattmann/git/grobid/grobid-home/schemas/rng/Grobid.rng\" schematypens=\"http://relaxng.org/ns
/structure/1.0\"?>\n<TEI xmlns=\"http://www.tei-c.org/ns/1.0\">\n\t<teiHeader xml:lang=\"en\"
>\n\t\t<fileDesc>\n\t\t\t<titleStmt>\n\t\t\t\t<title level=\"a\" type=\"main\">A Software Architecture-Based
Framework for Highly Distributed and Data Intensive Scientific Applications</title>..snip..</TEI>\n",
        "grobid:header_Title": "A Software Architecture-Based Framework for Highly Distributed and Data
Intensive Scientific Applications",
        "meta:author": "End User Computing Services",
        "meta:creation-date": "2006-02-15T21:13:58Z",
        "meta:save-date": "2006-02-15T21:16:01Z",
        "modified": "2006-02-15T21:16:01Z",
        "pdf:PDFVersion": "1.4",
        "pdf:encrypted": "false",
        "producer": "Acrobat Distiller 6.0 (Windows)",
        "resourceName": "ICSE06.pdf",
```

```
        "title": "Proceedings Template - WORD",
        "xmp:CreatorTool": "Acrobat PDFMaker 6.0 for Word",
        "xmpTPg:NPages": "10"
    }
]
```

Running GROBID using Tika-App CLI

Grab the latest version (2.8.0 or later) of the Tika-App and start the Grobid service by following the commands below.

You can run GROBID via Tika app with the following command on a sample PDF file.

```
java -cp grobidparser-resources:tika-app-2.8.0.jar:tika-parser-nlp-package-2.8.0.jar org.apache.tika.cli.
TikaCLI --config=grobidparser-resources/tika-config.xml -J PATH_TO_YOUR_PDF_FILE
```

Which should produce as output (e.g., if piped to "`` `python -m json.tool` ``" for pretty printing):

```
[
    {
        "Author": "End User Computing Services",
        "Company": "ACM",
        "Content-Length": "200435",
        "Content-Type": "application/pdf",
        "Creation-Date": "2006-02-15T21:13:58Z",
        "Last-Modified": "2006-02-15T21:16:01Z",
        "Last-Save-Date": "2006-02-15T21:16:01Z",
        "SourceModified": "D:20060215211344",
        "X-Parsed-By": [
            "org.apache.tika.parser.CompositeParser",
            "org.apache.tika.parser.journal.JournalParser"
        ],
        "X-TIKA:content": "<html xmlns=\"http://www.w3.org/1999/xhtml\">\n<head>\n<meta name=\"
access_permission:extract_for_accessibility\" content=\"true\" />\n<meta name=\"meta:save-date\" content=\"2006-
02-15T21:16:01Z\" />\n<meta name=\"grobid:header_Affiliation\" content=\"1 Jet Propulsion Laboratory California
Institute of Technology; 2 Computer Science Department University of Southern California\" />\n<meta name=\"
Content-Length\" content=\"200435\" />\n<meta name=\"dcterms:created\" content=\"2006-02-15T21:13:58Z\"
/>\n<meta name=\"Author\" content=\"End User Computing Services\" />\n<meta name=\"date\" content=\"2006-02-
15T21:16:01Z\" />\n<meta name=\"access_permission:can_modify\" content=\"true\" />\n<meta name=\"creator\"
content=\"End User Computing Services\" />\n<meta name=\"access_permission:modify_annotations\" content=\"
true\" />\n<meta name=\"Creation-Date\" content=\"2006-02-15T21:13:58Z\" />\n<meta name=\"grobid:
header_Address\" content=\"Pasadena, CA 91109 USA Los Angeles, CA 90089 USA \" />\n<meta name=\"meta:author\"
content=\"End User Computing Services\" />\n<meta name=\"created\" content=\"Wed Feb 15 13:13:58 PST 2006\"
/>\n<meta name=\"access_permission:fill_in_form\" content=\"true\" />\n<meta name=\"grobid:
header_FullAffiliations\" content=\"[Affiliation {orgName=Jet Propulsion Laboratory California Institute of
Technology , address=Pasadena, CA 91109 USA},Affiliation {orgName=Computer Science Department University of
Southern California , address=Los Angeles, CA 90089 USA}[Affiliation {orgName=Jet Propulsion Laboratory
California Institute of Technology , address=Pasadena, CA 91109 USA},Affiliation {orgName=Computer Science
Department University of Southern California , address=Los Angeles, CA 90089 USA}]\" />\n<meta name=\"grobid:
header_Class\" content=\"org.apache.tika.metadata.Metadata\" />\n<meta name=\"dc:format\" content=\"application
/pdf; version=1.4\" />\n<meta name=\"access_permission:can_print\" content=\"true\" />\n<meta name=\"Company\"
content=\"ACM\" />\n<meta name=\"xmp:CreatorTool\" content=\"Acrobat PDFMaker 6.0 for Word\" />\n<meta name=\"
resourceName\" content=\"ICSE06.pdf\" />\n<meta name=\"Last-Save-Date\" content=\..snip",
        "X-TIKA:parse_time_millis": "4302",
        "access_permission:assemble_document": "true",
        "access_permission:can_modify": "true",
        "access_permission:can_print": "true",
        "access_permission:can_print_degraded": "true",
        "access_permission:extract_content": "true",
        "access_permission:extract_for_accessibility": "true",
        "access_permission:fill_in_form": "true",
        "access_permission:modify_annotations": "true",
        "created": "Wed Feb 15 13:13:58 PST 2006",
        "creator": "End User Computing Services",
        "date": "2006-02-15T21:16:01Z",
        "dc:creator": "End User Computing Services",
        "dc:format": "application/pdf; version=1.4",
```

```
        "dc:title": "Proceedings Template - WORD",
        "dcterms:created": "2006-02-15T21:13:58Z",
        "dcterms:modified": "2006-02-15T21:16:01Z",
        "grobid:header_Address": "Pasadena, CA 91109 USA Los Angeles, CA 90089 USA ",
        "grobid:header_Affiliation": "1 Jet Propulsion Laboratory California Institute of Technology; 2
Computer Science Department University of Southern California",
        "grobid:header_Authors": "Chris A Mattmann 1,2 Daniel J Crichton 1 Nenad  Medvidovic 2 Steve  Hughes 1
",
        "grobid:header_Class": "org.apache.tika.metadata.Metadata",
        "grobid:header_FullAffiliations": "[Affiliation {orgName=Jet Propulsion Laboratory California Institute
of Technology , address=Pasadena, CA 91109 USA},Affiliation {orgName=Computer Science Department University of
Southern California , address=Los Angeles, CA 90089 USA}[Affiliation {orgName=Jet Propulsion Laboratory
California Institute of Technology , address=Pasadena, CA 91109 USA},Affiliation {orgName=Computer Science
Department University of Southern California , address=Los Angeles, CA 90089 USA}]",
        "grobid:header_Keyword": "\"D2 Software Engineering, D211 Domain Specific Architectures\"",
        "grobid:header_TEIJSONSource": "{\"TEI\":{\"text\":{..snip",
        "grobid:header_TEIXMLSource": "<?xml version=\"1.0\" encoding=\"UTF-8\"?>\n<?xml-model href=\"
file:///Users/mattmann/git/grobid/grobid-home/schemas/rng/Grobid.rng\" schematypens=\"http://relaxng.org/ns
/structure/1.0\"?>\n<TEI xmlns=\"http://www.tei-c.org/ns/1.0\">\n\t<teiHeader xml:lang=\"en\"
>\n\t\t<fileDesc>\n\t\t\t<titleStmt>\n\t\t\t\t<title level=\"a\" type=\"main\">A Software Architecture-Based
Framework for Highly Distributed and Data Intensive Scientific Applications</title>\n\t\t\t<
/titleStmt>\n\t\t\t<publicationStmt>\..snip</TEI>\n",
        "grobid:header_Title": "A Software Architecture-Based Framework for Highly Distributed and Data
Intensive Scientific Applications",
        "meta:author": "End User Computing Services",
        "meta:creation-date": "2006-02-15T21:13:58Z",
        "meta:save-date": "2006-02-15T21:16:01Z",
        "modified": "2006-02-15T21:16:01Z",
        "pdf:PDFVersion": "1.4",
        "pdf:encrypted": "false",
        "producer": "Acrobat Distiller 6.0 (Windows)",
        "resourceName": "ICSE06.pdf",
        "title": "Proceedings Template - WORD",
        "xmp:CreatorTool": "Acrobat PDFMaker 6.0 for Word",
        "xmpTPg:NPages": "10"
    }
]
```