

Documentation for the  
**JULIE Lab Medline Reader**

Matthias Mühlhausen

Jena University Language & Information Engineering (JULIE) Lab

Fürstengraben 30

D-07743 Jena, Germany

`matthias.muehlhausen@uni-jena.de`

## 1 Objective

The JULIE Lab Medline Reader is an UIMA Collection Reader that reads a set of XML documents that conform to the Medline PubMed XML format. It provides several document meta annotations like *author* or *publication date* and many more and can be used in a pipeline of UIMA components to initialize input objects (called CASes) for the subsequent annotators. It is part of the JULIE Lab NLP tool suite<sup>1</sup> which contains several UIMA components (NLP components from sentence splitting to named entity recognition), UIMA Collection Reader, UIMA Cas Consumer as well as a comprehensive UIMA type system.

## 2 Requirements and Dependencies

This Collection Reader is written in Java (*Java 6* or above is recommended - it should also run with Java 5 but is not tested with it) using *Apache UIMA version 2.2.1-incubation*<sup>2</sup>. It is developed using Linux, but it should also run on other platforms.

The input document format is *Medline/PubMed XML*.

The output takes place by annotation objects. The classes corresponding to these objects are part of a *JULIE Lab UIMA Type System* in the current version 2.1.<sup>3</sup>

---

<sup>1</sup>[www.julielab.de](http://www.julielab.de)

<sup>2</sup>[incubator.apache.org/uima](http://incubator.apache.org/uima)

<sup>3</sup>The *JULIE UIMA type system* can be separately obtained from [www.julielab.de](http://www.julielab.de)

## 3 Using the Medline Reader – Descriptor Configuration

In UIMA, each component is configured by a descriptor in XML. If you are not familiar with UIMA component descriptors please refer to the *UIMA SDK User's Guide*<sup>4</sup>, especially chapter 12 *Component Descriptor Editor User's Guide*.

A pre-configured descriptor is already contained in this package (see `desc/MedlineReader.xml`).

There are two parameters that can be configured in the Medline Reader descriptor:

**Input directory** As the Medline Reader reads PubMed XML, this documents must be provided in a folder of the file system. The parameter name is *InputDirectory*. The value is the path to a folder that contains the XML documents. All files in this folder are read by the Medline Reader, so it should only contain files of the right input format i.e. PubMed XML. Normally this is the only parameter defined in your descriptor.

**Single input file** It is also possible to configure the Medline Reader to read a one certain file. The parameter name is *InputFile*. The value of this parameter is a file name including a path name. This parameter is dominant, which means that if this parameter is declared, the *InputDirectory* parameter will be ignored.

## 4 Copyright and License

This software is Copyright (C) 2008 Jena University Language & Information Engineering Lab (Friedrich-Schiller University Jena, Germany), and is licensed under the terms of the Common Public License, Version 1.0 or (at your option) any subsequent version.

The license is approved by the Open Source Initiative, and is available from their website at [www.opensource.org](http://www.opensource.org).

## 5 Note

If you find bugs in the software or this documentation, or you if you have suggestions, please contact me. You are very welcome: [matthias.muehlhausen@uni-jena.de](mailto:matthias.muehlhausen@uni-jena.de)

---

<sup>4</sup>[incubator.apache.org/uima](http://incubator.apache.org/uima)