# WikiTailor

Technical Manual
Version 1.0

Cristina España-Bonet, Alberto Barrón-Cedeño and Josu Boldoba
WORKING DOCUMENT

February 14, 2016

**Abstract**

WikiTailor is a Java toolkit designed to extract and analise corpora from Wikipedia in any language and domain[1]. This document describes the prerequisites and the usage of the toolkit, the main characteristics of the systems involved in the extraction of the corpora, and some methods to evaluate the extractions.

[1]This work has been partially funded by the TACARDI project (TIN2012-38523-C02) of the Spanish Ministerio de Economía y Competitividad (MEC). quieres incluir al QCRI?

# Contents

# 1 Introduction

# 2 Prerequisites and Installation

This software is distributed as a stand-alone jar[2], so nothing else but a Java Virtual Machine is needed. The source code is also available[3] in which case, it can be compiled using the included Ant buildfile (see Section 2.3). In both cases a Wikipedia dump of the desired languages must be stored in a database to retrieve the articles off-line. Sections 2.1 and 2.2 describe how to download the data and create the database in a format compatible with WikiTailor.

## 2.1 Downloading the Dumps

In order to preprocess the Wikipedia articles a dump of the desired language(s) must be downloaded from the Wikimedia webpage:

`http://dumps.wikimedia.org/[LAN]wiki/[DATE]`

where, and from here on, `LAN` is a two-character language identifier with the ISO 639 codes[4] and `DATE` is the date of the dump in the `YEARMONTHDAY` format.

The dump consists of several sql tables. WikiTailor needs the general purpose tables:

`[LAN]wiki-[DATE]-pages-articles.xml.bz2`
`[LAN]wiki-[DATE]-pagelinks.sql.gz`
`[LAN]wiki-[DATE]-categorylinks.sql.gz`,

and three additional tables to relate articles in different languages and follow redirections:

`[LAN]wiki-[DATE]-langlinks.sql.gz`
`[LAN]wiki-[DATE]-page.sql.gz`
`[LAN]wiki-[DATE]-redirect.sql.gz`

The first group can be preprocessed and upload into a MySQL database using the JWPL DataMachine; the second group can be uploaded directly.

## 2.2 The JWPL DataMachine and the Database

JWPL (Java Wikipedia Library) is a Java-based application programming interface that allows to access all information in Wikipedia [**?**]. WikiTailor uses its DataMachine to create a MySQL database with a preprocessed Wikipedia dump and as a external library to parse the MediaWiki syntax.

The full process to deal with the dumps can be summarised in the following steps:

1. Download the JWPL DataMachine and the tables.sql file from `https://dkpro.github.io/dkpro-jwpl/`

2. Database setup. Create a MySQL database (use UTF-8 encoding) and afterwards the tables:

```
CREATE DATABASE IF NOT EXISTS [DBname] CHARACTER SET utf8 \
COLLATE utf8_general_ci

mysql -u [USER] -p [DBname] < tables.sql
```

[2] `http://cristinae.github.io/WikiTailor/dwnld/wikiTailor-v1.0.0.with-dependencies.tar.gz`
[3] `https://github.com/cristinae/WikiTailor`
[4] `http://www.loc.gov/standards/iso639-2/php/code_list.php`. ISO 639-1 is used when available.

where a DBname in the format [DBrootname]_wiki_[LAN]_[YEAR] is expected from WikiTailor.

3. JWPL preprocessing. Run the transformation:

   ```
   java -jar -Xmx4g datamachine.jar [LANGUAGE] \
   [CATEGORY:MAIN_CATEGORY_NAME] \
   [CATEGORY:DISAMBIGUATION_CATEGORY_NAME] [SOURCE_DIRECTORY]
   ```

   where:
   LANGUAGE is a string matching one in languages.txt in this release,
   MAIN_CATEGORY_NAME is the name of the main (top) category of the Wikipedia category hierarchy[5],
   DISAMBIGUATION_CATEGORY_NAME is the name of the category that contains the disambiguation categories[6] and,
   SOURCE_DIRECTORY is the path to the directory containing the data dumps.

   This should create a lot of new data files in a "output" subfolder of each input folder.

   Mind that the names of the main category or the category marking disambiguation pages may change over time. E.g. the English category for disambiguation pages was called "Disambiguation" for a long time, while now it is "All_disambiguation_pages".

   Examples:

   ```
   java -jar -Xmx4g \
   de.tudarmstadt.ukp.wikipedia.datamachine-1.0-jar-with-dependencies.jar\
   english Category:Contents Category:All_disambiguation_pages ./en/
   ```

   ```
   java -jar -Xmx2g \
   de.tudarmstadt.ukp.wikipedia.datamachine-1.0-jar-with-dependencies.jar\
   catalan Categoria:Principal Categoria:Pàgines_de_desambiguació ./ca/
   ```

   ```
   java -jar -Xmx2g \
   de.tudarmstadt.ukp.wikipedia.datamachine-1.0-jar-with-dependencies.jar\
   arabic تصنيف:محتويات_ويكيبيديا تصنيف:صفحات_توضيح ./ar/
   ```

4. Import the generated data files into de database.

   ```
   mysqlimport -u[USER] -h [HOST] -p -local \
   -default-character-set=utf8 [DBname] *.txt
   ```

A MySQL database can host the tables related to the interlanguage links, the langlinks. Just modify the names of the tables in the downloaded files to fit the WikiTailor nomenclature and upload them into the database.

1. Create the database:

[5] A complete list with the labels for the main category can be obtained in https://www.wikidata.org/wiki/Q1281.

[6] A complete list with the labels for the disambiguation pages can be obtained in https://www.wikidata.org/wiki/Q1982926.

```
CREATE DATABASE IF NOT EXISTS [DBrootname_pairs] \
CHARACTER SET utf8 COLLATE utf8_general_ci
```

2. Within the sql file, change the name of the table ‘`langlinks`‘ into ‘`wiki[LAN]_[YEAR]_langlinks`‘. Do the equivalent modification for ‘`page`‘ and ‘`redirect`‘.

3. Upload the tables:

```
mysql -u[USER] -p [DBrootname_pairs] < [LAN]wiki-[DATE]-langlinks.modified.sql
mysql -u[USER] -p [DBrootname_pairs] < [LAN]wiki-[DATE]-page.modified.sql
mysql -u[USER] -p [DBrootname_pairs] < [LAN]wiki-[DATE]-redirect.modified.sql
```

### 2.3 WikiTailor Installation

## 3 Description and Usage

### 3.1 Extraction of Content Articles

### 3.2 xxx

#### 3.2.1 Including a new Language

/home/cristinae/pln/workspace/lump2/lump2-aq-textextraction/HOWTO.txt

## 4 Evaluating the Extractions

## 5 Additional Utilities

## 6 Towards WikiParalel