

# Instruction

version: 2.2

This instruction will guide you through every step on using the validation tools for validating a submission run. The validation tool package includes three programs: CrosslinkValidation (GUI), RunChecker(CMD), XML2TXT(CMD).

## Contents

1	CrosslinkValidation (GUI) .....	1
2	RunChecker .....	4
3	XML2TXT .....	5


## 1 CrosslinkValidation (GUI)

### 1.1. Lauching the program

- On Unix-like system (e.g. Linux)  
\$sh run.sh
- On Windows  
Double click on the run.bat or execute it in the command line console.

### Specification of the Directory of the CJK Wikipedia Collection.

- For the first time using this tool, you must specify the directory of the Wikipedia collections by using the menu bar function **Corpus** under **Utility**.
- The directory or the home of the CJK Wikipaida corpora will be the directory holding the **zh** , or **ja** , or **ko** folder, or all above folders unzipped from the downloaded Wikipedia collection. For example,



Location: /data/corpus/wikipedia/CJK/xml-v2

Name	Size	Type
ja	1 item	folder
ko	1 item	folder
zh	1 item	folder
ja-pages.tar.bz2	1.1 GB	Tar archive (bzip-compressed)
ko-pages.tar.bz2	163.1 MB	Tar archive (bzip-compressed)
zh-pages.tar.bz2	380.6 MB	Tar archive (bzip-compressed)

Figure 1. the collections home

So, /data/corpus/wikipedia/CJK/xml-v2 is the home directory of the CJK Wikipedia corpora.

- Once the directory of the collections have been set up correctly, the tool will show the current Wikipedia collection home directory (see Figure 6). You do not need to specify them again, except the directory have been changed.

1) Please place the Wikipedia DTD file - **article.dtd** just in the parent directory of the package.

- As we can see each Wikipedia document has been specified the DTD, `<!DOCTYPE article SYSTEM "../article.dtd">`, the DTD file must be placed in the right location for the Wikipedia document. This can be ignored if the program is started using the “run.bat” or “run.sh” script file. If any error happens and indicates missing the DTD file, please make the article.dtd file is located in the right place.

2) **Split a submission:**

- In order to validate your submission and see how the specified anchor offset and length match up with the text in the topic file, each submission needs to be split into multiple run files according to topic first.
- To do so, select the function **Split Run** under **Utility**, then locate the submission file and click Ok.
- Information will be given for the location of the split files.

3) **Sample Wikipedia submission runs:**

- After splitting the run, you should select the function **Load** under **Utility** to load a submission. Once the loading is success, the tool will display all content you need to check the submission run.
- One sample submission XML file, **SAMPLE\_A2F\_RUN\_01.xml**, have been prepared in the directory, **Sample\_Submission**, for this instruction. You can split it then use **Load** function to browse one file at a time.
- If the loading was not success or the tool did not display the content correctly, it means there are something going wrong with the submission XML format.

4) **Topic Pane, Link Pane, Table Pane interaction:**

- Please refer to Figure 8 to 11 regarding to the area of each pane and the interaction scheme.
- You can click on an Anchor Text or Best Entry Point icon on the Topic Pane to just to the 1<sup>st</sup> respective outgoing target link.
- On the Link Pane, you can simply either right click to go to the *Next* link or left click to go back the *Previous* link.
- The Table pane offers a convenient means for the user to navigate to any of the topic and the target link by just clicking on a Table Cell. This Table also show the Offset and Length for each Anchor and BEP.

5) **Outgoing and Incoming Links:**

- There are 2 menu item functions, **Outgoing** and **Incoming**, are located under the menu function, **Linking**.
- Incoming mode has been disabled because only outgoing links need to be specified in NTCIR 9 crosslink task.

6) **The detail interaction scheme is described below.**

## 1.2. The Detailed Description

- For the first time, the tool will ask you to specify the home of your Wikipedia collection (see Figure 1).
- Click **OK** to let the tool launch entirely
- From the Menu bar, there is a menu item called **Utility**. Click on it and select the option, **Corpus**.
- As you can see, the CJK Wikipedia Collections Dialog is open and you can browse and select the directory where the Wikipedia collections were located.
- Once you have specified the directory of the collections, the tool will remember it and show it to you every time you launch the tool. The directory only needs to be changed when the collections have been moved.
- Next, you need to do is to split a submission file if you haven't do so, then load a single run file for a particular topic. From the Menu bar, there is a menu item called **Utility**. Click on it and select the option, **Load**. A Window file browser will open for you to select a topic run file. One at a time, which means you can only load a submission file for only one topic and display it each time.
- You can click on either an *Anchor Text* or *BEP Icon* on the Topic pane to display its relative links showing on the Link Pane.
- Right-Click* on the Link Pane will move to the Next link while *Left-Click* on the Link Pane will be one link backward.
- You can also click on a Table Cell on the RHS Pane to jump into that particular link. According to the **Row** you click, the tool will display the corresponding topic and link on the respective pane.
- The general information (e.g. *topic title*, *anchor name* and *target document title*) will be showed on the top banner just below the menu bar. From the RHS Table Pane, you can see the *Offset* and *Length* of each anchor

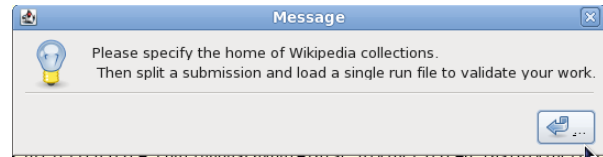


Figure 2. Empty directory of the collections

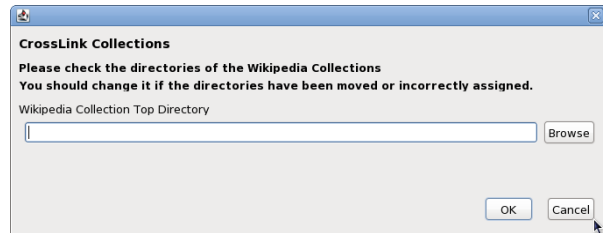


Figure 3. Specification of home of the CJK Wikipedia collection

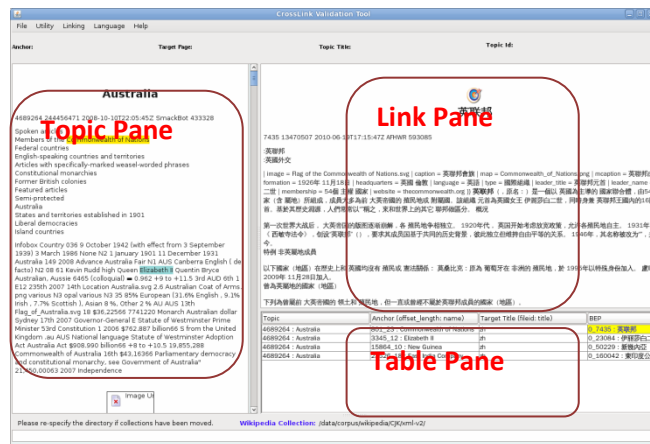


Figure 4. The crosslink submission

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <!-- generated by CLiX/Wiki2XML [MPI-Inf, MMCI@UdS] $LastC
3 <!DOCTYPE article SYSTEM "../article.dtd">
4 <article xmlns:xlink="http://www.w3.org/1999/xlink">
5 <relationship confidence="0.8" wordnetid="113780719">
6 <header>
7 <title>Online dating service</title>
8 <id>169553</id>
9 </revision>

```

Figure 5. The Wikipedia Page DTD Information

and BEP. According to this information, you can adjust your system to correct the position of each text anchor and best entry point. A general idea is that although an exact “hit” of position (i.e. offset and length) is perfect, these anchors and BEPs are NOT expected to match exactly their respective position on the document but nearby or overlapping.

martial training which began in childhood. Some First Nations men, and more rarely some women, were called warriors only after they had

Please re-specify the directories if collections have been moved.

Wikipedia Collection: /data/corpus/wikipedia/CJK/xml-v2/

**Figure 6. The CJK Wikipedia collections directory message**

## 2 RunChecker

This script will help you to check if you have all anchors correctly specified with the given offset and length. If recommended anchors have correct offset and length, information will be directed to stdout, otherwise the detailed error message will be redirected to stderr.

### Usage:

- On Unix-like system (e.g. Linux)  
`./checkrun.sh run1.xml [run2.xml] ...`
- On Windows  
`> checkrun.bat run1.xml [run2.xml] ....`

For example, showing only the incorrect anchor information is desired, use:

```
./checkrun.sh run1.xml 1>/dev/null
```

or

```
checkrun.bat run1.xml 1> NUL
```

### 3 XML2TXT

This tool helps return the text in the input file for the given offset and length, or replace all XML tags with spaces.

#### Usage:

- On Unix-like system (e.g. Linux)  
`./xml2txt.sh [-o:offset:length] input_xml`  
return the text with the given offset and length.

Or

`./xml2txt.sh input_xml`  
remove all the tags, output to stdout.

- On Windows

`./xml2txt [-o:offset:length] input_xml`  
return the text with the given offset and length.

Or

`./xml2txt input_xml`  
remove all the tags, output to stdout.