DBLP data retriever IEEE explorer and ACM crawler.

USER'S MANUAL - *Lombrico*

Written by Sergio Andaloro and Gioele Salvatore Antoci, Politecnico di Milano – 2011.

Index

1. **Brief description**

This program has been written to parse all the papers in a conference reported by IEEE or ACM associations and to write the output as an HTML page that will respect rules imposed by DBLP standards (http://www.informatik.uni-trier.de/~ley/db/about/faqformat.html). This program recognizes the papers of a conference looking for some kind of tags and so only the pages that will match the pattern of the IEEE and ACM conference pages of today will have a regular output with a normal execution. Later we will explain the patterns that we found in the pages in object. In order to prove the functionalities of our program we used a sample of 1000 web page for both IEEE and ACM standard page.

Requirements:
-libcurl3 version 7.21.0 or greater.
-libpcre3 version 8.02-1 or greater.

2. **How to use**

To run *lombrico* write *./lombrico + "URL of conference"* in the terminal. *Lombrico* warns you that files in output folder will be overwritten (without deleting older ones). If the URL is from IEEE domain then choose if a paper without author(s) have to be inserted in html output.

In output folder, called *outputForDblp*, you can find html pages that respect the rules of DBLP standard.

3. IEEE HTML standard page

We based our studies on a IEEE page like the following one:
http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5507538

The code we write is made on the standard HTML page of the IEEE conference where the HTML that we used for parsing is formatted as follow:

```
<!-- CODE FOR CHECKING ABSTRACTPLUS FOR USERS WHO HAVE VIRTUAL JOURNAL PACKAGES -->

                    <li class="noAbstract">
                        <div class="header">
                            <div class="select">
<input name="articlesCheckBox" type="checkbox" value="274634" id="274634" />
                            </div>
                            <div class="detail">
                                <h3>
                <a href='/xpls/abs_all.jsp?arnumber=274634'>Earthing, protection,
  miscellaneous [of railway power systems]</a>
                                </h3>

                                <b>Courtois, C. </b>
                                <br>

                                <b>Page(s):</b>

                                    4/1 - 4/3

                <p class="links"><a href="/xpls/abs_all.jsp?arnumber=274634">Abstract</a>

| Full Text: <a href="/stamp/stamp.jsp?tp=&arnumber=274634">PDF</a> (104KB)


                                </p>

                            </div>
                        .
                        .
                        .
</div><!-- END MAIN CONTENT -->
```

In orange we found the first string that matches the beginning of each paper. In brown the start and the end of the matching for the title. From the green tag, between and there's the author. We look for the author's tags until the first red tag. Then, in the starting from the first red tag to the second one ("<") there is the pages string. If it's available an electronic edition of the paper in pink we can

see the tag used to found its URL. The string in light blue indicates the end of the HTML concerning to the papers.

If in the main HTML page there is more than one page there will be something like that:

```
<li>
<span>1</span>
</li>

<li><a href="#" onclick="javascript:gotoPage('2')">2</a></li>

<li><a href="#" onclick="javascript:gotoPage('3')">3</a></li>
```

So the program will automatically download each page and create a multipage HTML.

4. ACM HTML standard page

We based our studies on a IEEE page like the following one:
http://portal.acm.org/citation.cfm?id=1455770&picked=prox

In the standart ACM page we look for the part where is sited the link for the table of contents. The link we search start with the first red string matching and end with the second red one.

```
ColdFusion.Bind.register([],{'bindTo':'prox','bindExpr':['tab_about.cfm?id=1080402&type=proceeding&parent_id=1080
402&parent_type=proceeding&title=Proceedings%20of%20the%202nd%20symposium%20on%20Applied%20perceptio
n%20in%20graphics%20and%20visualization&toctitle=ACM%20Symposium%20on%20Applied%20Perception%20in%20
Graphics%20and%20Visualization%202005&tocissue_date=&notoc=0&usebody=tabbody&tocnext_id=&tocnext_str=&t
ocprev_id=&tocprev_str=&toctype=conference&cfid=35954070&cftoken=76285039']},ColdFusion.Bind.urlBindHandler,
true);
```

After that parsing the table of content is downloaded and analyzed. A standard part of the HTML of a table of contents page look like the following one:

```
<td colspan="1"><span style="padding-left:20"><a
href="citation.cfm?id=1455772&CFID=35954070&CFTOKEN=76285039">The good, the bad, and the provable</
a></span></td> </tr> <tr><td></td> <td>
<span style="padding-left:20">
<a href="author_page.cfm?id=81100547147&CFID=35954070&CFTOKEN=76285039">Mart&#237;n
Abadi</a>
</span>
</td> </tr> <tr> <td></td>
<td> <span style="padding-left:20">Pages: 1-1</span></td>
</tr> <tr><td></td>
<td> <span style="padding-left:20">doi&gt;<a href="http://dx.doi.org/10.1145/1455770.1455772"
title="DOI">10.1145/1455770.1455772</a></span></td>
</tr> <tr><td></td><td>
<span style="padding-left:20">
Full text: <a name="FullTextPdf" title="FullText Pdf"
href="ft_gateway.cfm?id=1455772&type=pdf&CFID=35954070&CFTOKEN=76285039" target="_blank"><img
src="imagetypes/pdf_logo.gif" alt="Pdf" class="fulltext_lnk" border="0" />Pdf</a>
</span>
</td>
```

With the red string we found the part after which there's the title, with the orange one the string after which we found the author(s) until the span tag in green that indicates the end of the section where we found the author(s). In blue there's the tag to match if we want to find the number of page for the paper and finally in pink the string "Full text:" (but also "Available formats: " and "Full text available:") that indicates the presence of an electronic resource.