



LogCLEF 2010

Giorgio Maria Di Nunzio

2010/05/17

v 1.0

Description of the The European Library (TEL) Action and HTTP Log Files



List of the files

For the LogCLEF 2010 Track the following files are available and can be downloaded from the following links (password protected)

January 2007 – June 2008

- Action log files
 - A postgresql table with all log records
<http://ims.dei.unipd.it/logclef/logclef2009.backup>
 - A zip file of a text file (semi-colon separated values) with the log records
<http://ims.dei.unipd.it/logclef/logclef.zip>
 - A 7zip (zipped) file of a text file (semi-colon separated values) with all the records
<http://ims.dei.unipd.it/logclef/logclef.7z>
- HTTP log files
 - 18 zip files. Each file contains one month of log data from January 2007 till June 2008
<http://ims.dei.unipd.it/logclef/http/ex0701.zip>
<http://ims.dei.unipd.it/logclef/http/ex0702.zip>
...(change URL to get the remaining months)
<http://ims.dei.unipd.it/logclef/http/ex0805.zip>
<http://ims.dei.unipd.it/logclef/http/ex0806.zip>

January 2009 – December 2009

- Action log files
 - A zip file of a csv file (comma separated values) with the log records
<http://ims.dei.unipd.it/logclef/logclef2.zip>

Additional info

- Collections
 - A csv file containing the description of the TEL collections
 - http://ims.dei.unipd.it/logclef/TEL_collections.csv



Description of the Search action logs

The *actionlog* table saved in the logclef2009.backup file has been created with a postgresql DBMS according to the following definition:

```
CREATE TABLE actionlog
(
  id bigint NOT NULL,
  userid character varying(25) NOT NULL,
  userip character varying(15) NOT NULL,
  sesid character varying(26) NOT NULL,
  lang character varying(3) NOT NULL,
  query character varying(250),
  "action" character varying(30),
  colid character varying(5),
  nrrecords integer,
  recordposition character varying(25),
  sboxid character varying(50),
  objurl character varying(250),
  date timestamp without time zone,
  CONSTRAINT actionlog_pkey PRIMARY KEY (id)
)
WITH (OIDS=FALSE);
ALTER TABLE actionlog OWNER TO postgres;
CREATE INDEX actionlog_sesid_idx
  ON actionlog
  USING btree
  (sesid);
```

Please, note that the original owner is the default postgres user.

The table contains a total of 1,866,330 records starting from the first of January 2007 until the 30th of June of 2008.

The meaning of each record is explained in the following section.



Description of the fields of the actionlog table

These are the fields of the actionlog table:

- id: identifier of the record, not necessarily contiguous numbers.
- userid: the identifier of the user; “guest” if the user did not log in, a number if the user logged in.
- userip: the IP address of the user, semi-obfuscated (first two bytes in clear, last two bytes obscured, for example 127.0.xxx.xxx or 127.0).
- sesid: PHP session id created on session start. it can be ‘null’ value (where null in this case is not an empty value, but a sequence of characters which forms the word ‘null’).
- lang: the language of the interface selected by the user in the portal
- query: the text of the query. There may be some unknown character encoding problems, or some inconsistencies in the Common Query Language. A possible query may have one of the following values:
 - (title all "keywords")
 - (creator all "keywords")
 - (subject all "keywords")
 - (type all "keywords")
 - (language all "keywords")
 - (isbn all "keywords")
 - (issn all "keywords")
 - ("keywords")

The combination of atomic queries is done by means of boolean operators, for example (title all "keywords") and (creator all "keywords") .

- action: the action performed by the user, a complete list of actions is presented below.
- colid: the identifier of the collection that has been involved by the action of the user.
- nrRecords: the number of records retrieved for the collection involved by the action of the user.
- recordPosition: position of the item in the total record list.
- sboxid: identifier for a remote searchbox which query the Web portal.
- objurl: the URL of the object being viewed.
- date: timestamp of the action yy-mm-dd hh:mm:ss.

The available actions are the following ones:

- search_sim: start search from simple search form.
- search_adv: start search from advanced search form
- search_res: start search from search form in the results page
- search_res_rec_any, search_res_rec_all: start search from a full record view by clicking on search(magnifying glass) icon in the record’s available fields
- search_url: start search from URL query string. This string may also have a domain name attached to it (search_url_www.domain.org) if it is coming from a remote tel search – minitel (a marketing tool)
- view_brief: display the short title – list of records



- view_full: display long title - individual record. Activated when a user clicks on a title link in the list of brief records displayed (20 per page), or when a user clicks on the previous or next link when already viewing a full record.
- jump_to_page: when displaying brief titles, a user can enter a numerical value for skipping several pages of records.
- available_at: "Available at Library" link clicked to view record in native national library interface.
- see_online: "See online" link clicked to see object in native interface.
- page_brief: user clicked on "next" or "previous" buttons to change record lists (20 per page).
- col_set_theme: collections chosen from theme list.
- col_set_theme_country: collections chosen from country list on homepage or results page.
- col_set_country: collections chosen from all collections tab (collections listed by country).
- col_set_subj: collections chosen from subject list.
- col_set_desc: collections chosen by searching by description.
- col_set_defaultCollections: default list reinstated.
- option_save_session_favorite: Session favorite saved.
- option_send_mail: Record sent by email.
- options_save_reference: Record saved for reference manager use.
- service_denmark: full record service link used.
- service_hungary: full record service link used.
- service_netherlands: full record service link used.
- service_uk: full record service link used.
- service_all: full record service link used.
- show_help_helpfilename: "help" link clicked.
-



Description of the CSV log files

The two compressed (zip and 7zip) files contain the same comma-separated-values file.

These files are text files where each row contains the description of a log record. The list of the fields is exactly the same as the list of fields which appear in the *actionlog* table description.

Each field is separated by a semi-colon (“;”) or by a comma (“,”).



Description of the HTTP logs

The HTTP log files are saved in 18 text files (zipped). The extension of each file is .cv, although the character which separates the fields of a record is a space (" ") and not a comma, or a semi-colon.

Each record contains the following fields:

- date: year-month-day.
- time: hour:minute:second.
- HTTP method: for example GET, HEAD, POST, etc.
- URI stem: the path of the requested file.
- URI query: the string of the query in the URL, if any. A "-" is recorded if the URI query is null.
- IP address: the address of the client, only the first two bytes are shown (e.g. 127.0).
- User agent: the user agent of the client.
- Cookie: the cookie sent to/by the client*.
- Referrer: the URL of the resource which linked the client to TEL.

The Cookie field is divided into subfields by semi-colons ";". The subfields are (some of them can be ignored for this study):

- cTargets: the identifiers of the collections selected by the user (or selected by default by TEL), each selected collection is recorded as a pair "collections:a0037".
- __utmX values: where "X" can vary. Ignore these fields.
- cTargetThemes: ignore this field.
- TELSESSID: the identifier of the session. It is the same identifier recorded in the action logs under the name "sesid". This is an important field if you want to cross analyze action logs and HTTP logs.