



**web
archiving**



hackmeeting 2009

Memoria e attivismo digitale

<http://www.archivio99.org>

@ Archivio storico “**Il sessantotto**”
centro studi politici e sociali
Firenze

<http://associazioni.comune.firenze.it/archivioilsessantotto>



404

not found



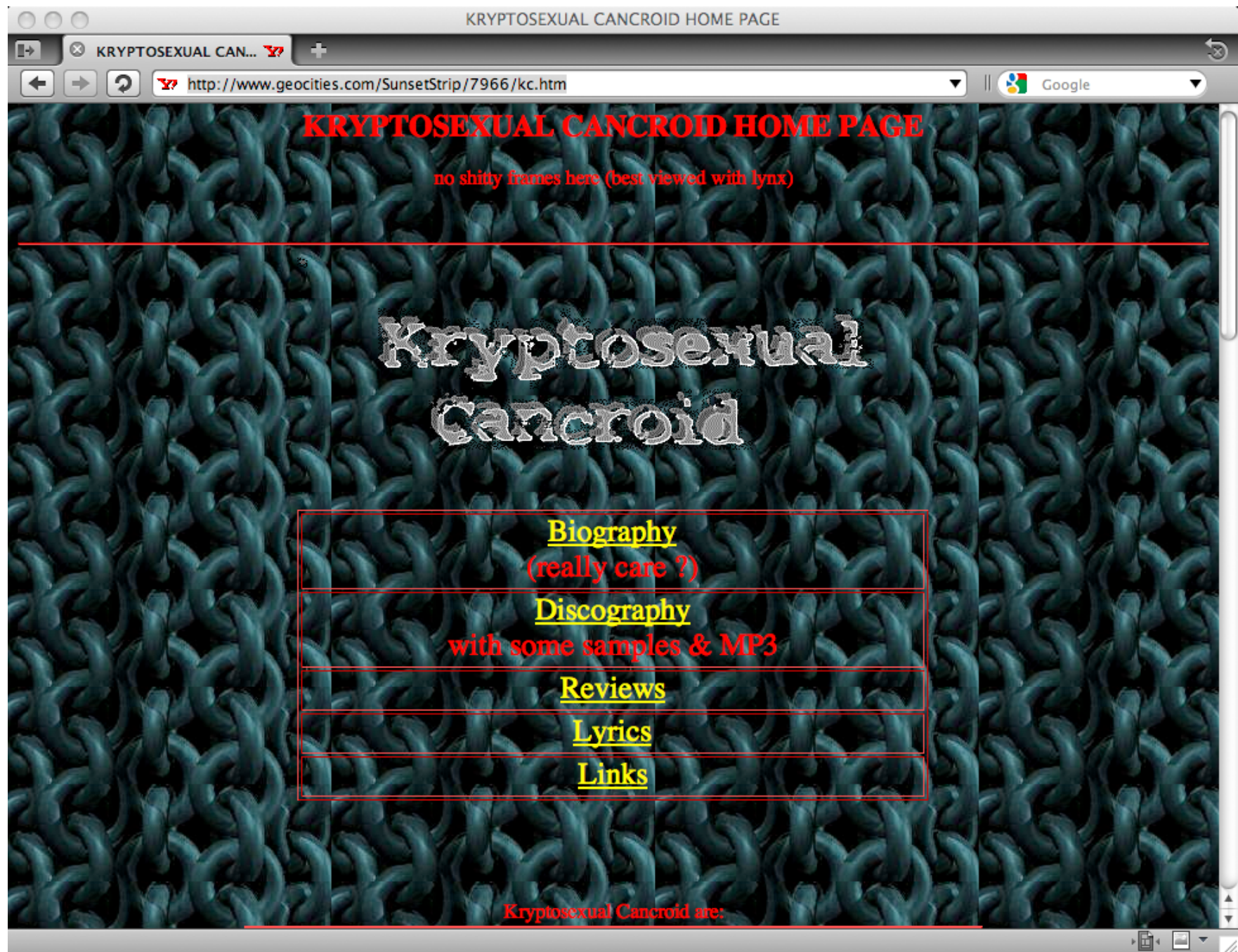
DATAPOCALYPSE

<http://ascii.textfiles.com/archives/1617>

31.10.2008 AOL Hometown

31.12.2009 Geocities





- ▶ il **webarchiving**, concetti generali
- ▶ **web crawlers**
 - ▶ browser, wget, httrack, web curator
 - ▶ heritrix
 - ▶ il formato WARC
 - ▶ problemi e limitazioni
- ▶ **dissemination**
 - ▶ wayback machine
 - ▶ nutchwax, search-tools
- ▶ **risorse e progetti**

web curation

http://en.wikipedia.org/wiki/Web_archiving

▶ selection

▶ **acquisition**

▶ validation

▶ **dissemination**

repository long
term preservation



selection

attività redazionale per la selezione dei seeds

ambito della scelta:

- ▶ site-centric
- ▶ topic-centric
- ▶ domain-centric



acquisition

- ▶ client-side archiving
- ▶ transactions archiving
- ▶ server-side archiving



client-side archiving

crawlers

software che salvano le pagine web e i relativi contenuti in files o specifici contenitori.

- ▶ browser
- ▶ zotero
- ▶ wget
- ▶ httrack
- ▶ harvestman
- ▶ web curator tool
- ▶ heritrix

archiviazioni minime con l'uso del browser

File → Save As

pro: semplice, immediato

cons: una sola pagina.

formati contenitori non standard

se si vogliono salvare i contenuti (immagini, etc)



webarchive <http://en.wikipedia.org/wiki/Webarchive>



kde war [http://en.wikipedia.org/wiki/KDE_WAR_\(file_format\)](http://en.wikipedia.org/wiki/KDE_WAR_(file_format))



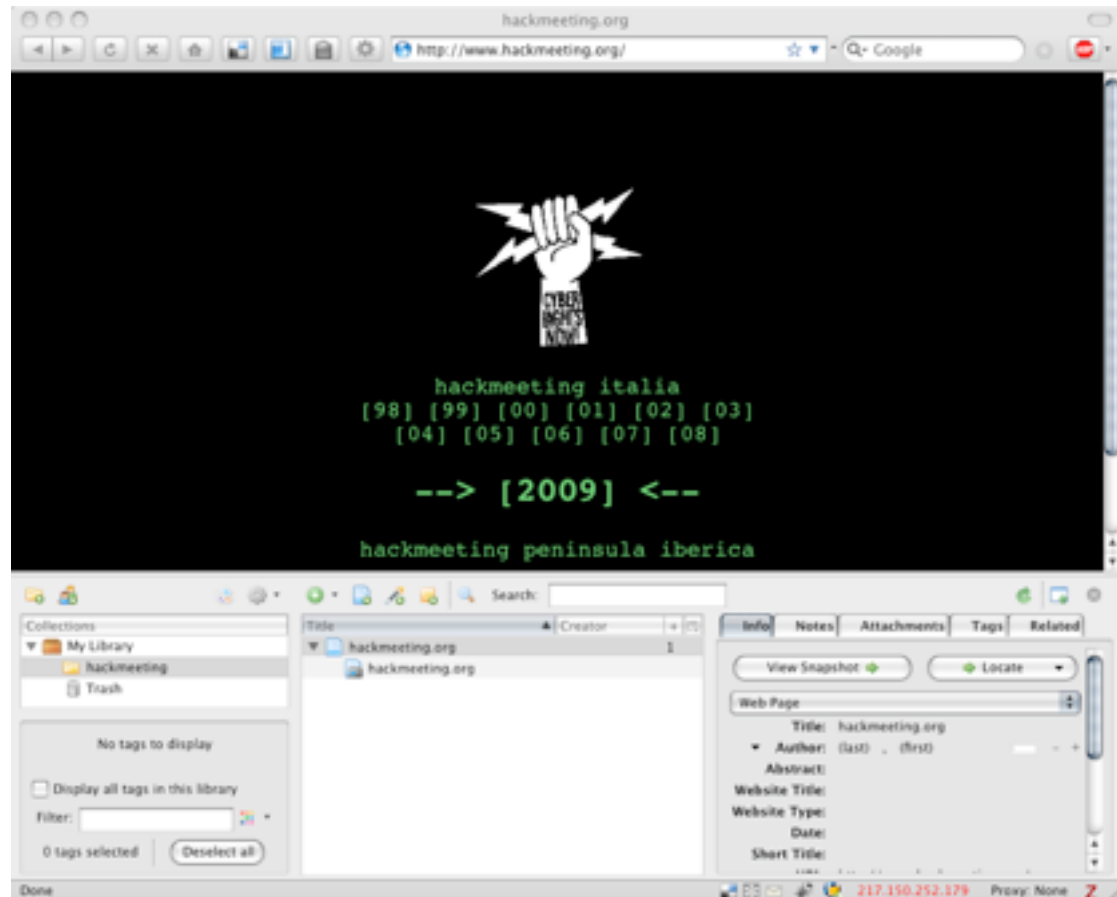
mhtml <http://en.wikipedia.org/wiki/MHTML>



file + directory

<http://www.zotero.org>

http://www.zotero.org/support/archive_the_web





gnu wget

<http://www.gnu.org/software/wget/>

```
$ wget -mEk http://www.example.com/site
```

-m, --mirror

shortcut for -N -r -l inf --no-remove-listing.

-E, --html-extension

save HTML documents with '.html' extension.

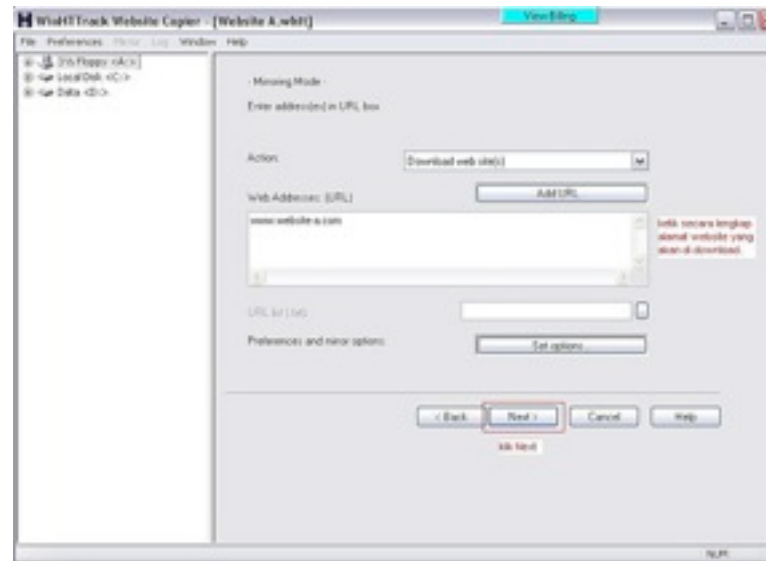
-k, --convert-links

make links in downloaded HTML point to local files.



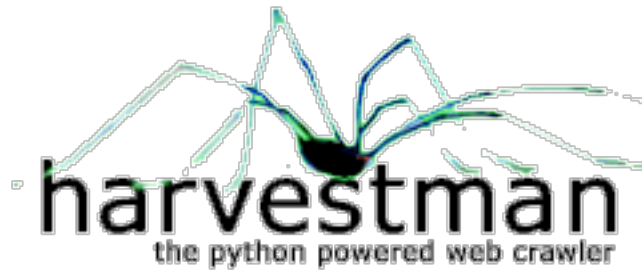
httrack

<http://www.httrack.com>



\$./httrack http://www.example.com





<http://harvestman.everythingability.com>

<http://code.google.com/p/harvestman-crawler/>





<http://webcurator.sourceforge.net>

Applicazione web (in java) per la gestione di
di harvesting web selettivo.





heritrix

<http://crawler.archive.org>

- ▶ java
 - ▶ affidabile per grosse catture
 - ▶ tecnologie standard
 - ▶ estensibile e distribuibile
 - ▶ largo impiego: archive.org
-
- ▶ documentazione per tecnici
 - ▶ interfaccia per tecnici

WARC

ISO 28500:2009

http://bibnum.bnf.fr/WARC/warc_ISO_DIS_28500.pdf

il formato di file **warc** (web archive) permette di concatenare record multipli (oggetti digitali) ognuno dei quali composto da un set minimo di headers (metadati) e da un arbitrario blocco di dati, in un unico file.

<http://code.google.com/p/warc-tools>

- supportato da heritrix
- plugin aggiuntivo per httrack
- in sviluppo patch per wget



WARC/1.0
WARC-Type: response
WARC-Target-URI: <http://www.archive.org/images/logoc.jpg>
WARC-Warcinfo-ID: <urn:uuid:d7ae5c10-e6b3-4d27-967d-34780c58ba39>

WARC-Date: 2009-06-01T17:20:24Z
WARC-Block-Digest: sha1:UZY6ND6CCHXETFVJD2MSS7ZENMWF7KQ2
WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
WARC-IP-Address: 207.241.233.58
WARC-Record-ID: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>
Content-Type: application/http;msgtype=response
WARC-Identified-Payload-Type: image/jpeg
Content-Length: 1902

HTTP/1.1 200 OK
Date: Tue, 01 Jun 2009 17:18:40 GMT
Server: Apache/2.0.54 (Ubuntu)
Last-Modified: Mon, 01 Jun 2009 22:28:51 GMT
ETag: "3e45-67e-2ed02ec0"
Accept-Ranges: bytes
Content-Length: 1662
Connection: close
Content-Type: image/jpeg

[image/jpeg binary data here]



problemi e limitazioni

- ▶ esclusione da robots.txt
- ▶ parti nascoste, il Deep Web
- ▶ risultati diversi a seconda dell'Agent
- ▶ trappole (i calendari, url infinite)
- ▶ ...

ecn.org

Crawl Name: **ecn-20080614115102**
Duration Time: **5d19h40m4s314ms**
URIs Crawled successfully: 152129
URIs Failed to Crawl: 1785
URIs Disregarded: 252
Processed docs/sec: 0.3
Bandwidth in Kbytes/sec: 7
Total Raw Data Size in Bytes: 3891984728 (3.6 GB)
Novel Bytes: 3892000348 (**3.6 GB**)

tmcrew.org

Crawl Name: tmcrew-20080623153450
Duration Time: **21h21m20s551ms**
URIs Crawled successfully: 17137
URIs Failed to Crawl: 29
URIs Disregarded: 11
Processed docs/sec: 0.22
Bandwidth in Kbytes/sec: 6
Total Raw Data Size in Bytes: 480200860 (458 MB)
Novel Bytes: 480239802 (**458 MB**)

strano.net

Crawl Name: **stranonet-20080614115102**
Duration Time: **8h17m2s422ms**
URIs Crawled successfully: 2919
URIs Failed to Crawl: 12
URIs Disregarded: 0
Processed docs/sec: 0.1
Bandwidth in Kbytes/sec: 3
Total Raw Data Size in Bytes: 108605680 (104 MB)
Novel Bytes: 108605680 (**104 MB**)



dissemination



wayback

<http://archive-access.sourceforge.net/projects/wayback/>

applicazione in java/tomcat

indicizza i contenuti di arc/warc e li rende disponibili
in http, preservando le url originali
e le revisioni nel tempo

http://web.archive.org/web/*/http://www.hackmeeting.org



search



<http://archive-access.sourceforge.net/projects/nutchwax>

search-tools

Open source web archive search tools

<http://code.google.com/p/search-tools>





<http://www.archiveteam.org>

deathwatch

geocities crawl



IWAW <http://iwaw.net>
International Web Archiving Workshop



www.archive.org
www.archive-it.org



european archive
<http://www.europarchive.org/>



<http://www.liwa-project.eu>

<http://www.hanzoarchives.com>



327M zone-digitali.iso

CD ROM - ZONE DIGITALI

1998

Isole Nella Rete - ECN
Tactical Media Crew
Strano Network
Avana
Banca dati della Memoria
Stragi di Stato
Decoder
Neural