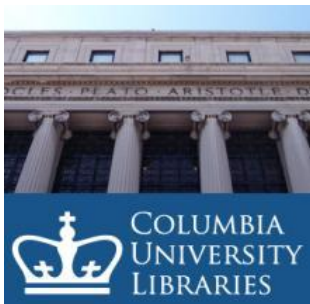


Solr-powered full-text and metadata search in the

HUMAN RIGHTS WEB ARCHIVE

CENTER FOR HUMAN RIGHTS DOCUMENTATION & RESEARCH AT COLUMBIA UNIVERSITY



IIPC General Assembly 2015
AWG

ALEX
THURMAN

ERIC
O'HANLON

HUMAN RIGHTS WEB ARCHIVE

CENTER FOR HUMAN RIGHTS DOCUMENTATION & RESEARCH AT COLUMBIA UNIVERSITY

The Human Rights Web Archive @ Columbia University is a searchable collection of archived copies of human rights websites created by non-governmental organizations, national human rights institutions, tribunals and individuals. Collecting began in 2008 and has been ongoing for active websites. New websites are added to the collection regularly.

The HRWA is an initiative of the [Center for Human Rights Documentation & Research](#) and is a key focus of the Columbia University Libraries' [Web Resources Collection Program](#). The HRWA was made possible by generous support from the Andrew W. Mellon Foundation. [Learn More](#) »

A selection of sites relating to the subject **Civil society**.



ABONG--Associação Brasileira de Orga...

[Site Details](#)



Centro Bonó

[Site Details](#)



Centro de Estudios Sociales Padre Jua...

[Site Details](#)



Human Rights Web Archive portal

Background

- Columbia web archiving began in 2008, first collection topic chosen was Human Rights (NGOs, NHRIs, blogs)
- Harvesting done with Archive-It
- Local access portal developed with Mellon funding 2011-2012
- Collection is re-crawled quarterly, with data downloaded from Archive-It and indexed locally
- Apart from ongoing data downloads, no further portal development since 2012 due to lack of dedicated funding for web archiving technical work

Human Rights Web Archive portal

Features

- CUL customized access portal to Human Rights Web Archive
- Browse archived websites by Title, URL, Subject, Place (Geographic focus), Language, using facet data pulled in from CLIO MARC records
- Search website descriptions (i.e. MARC metadata)
- Search page full text (based on local Solr indexing of .WARC file data downloaded from Archive-It and extracted into a mysql database)
- "Nominate a site" forms for site owners and general public
- Search extension -- search query extended (by user's opt-in) via API to any of a list of external resources related to web archives and/or human rights
- Search expansion -- If query term matches any term on pre-constructed list of names of people groups and their synonyms, users offered the option of expanding their search to include the relevant synonyms

FEATURED

WEBSITES

URLS

SUBJECTS

PLACES

LANGUAGES

Browse websites by title.

A-Z

Z-A



"İnsan Hüquqları XXI əsr-
Azərbaycan" Fondu

5.18 Kinyöm Chaedan

ABONG--Associação
Brasileira de Organizações
Não Governamentais

Abuelas de Plaza de Mayo

Access to Justice

Action Group for Health,
Human Rights and HIV/AIDS
(AGHA)

Ação Brasileira pela Nutrição
e Direitos Humanos

Adalah

Addameer Prisoner Support
and Human Rights
Association

Advocacy Forum--Nepal

droits de l'homme en Tunisie

CSVR, Centre for the Study
of Violence and
Reconciliation

Cultural Survival

The Cyprus Action Network

Danish Institute of Human
Rights

Darfur Consortium

Defensor del Pueblo de la
Nación

Defensor del Pueblo [Spain]

Defensores en linea.com

La Defensoria de los
Habitantes

La Defensoria del Pueblo,
Estado Plurinacional de
Bolivia

al-Majlis al-Istishārī li-Ḥuqūq
al-Insān

al-Majlis al-Waṭanī lil-Ḥurrīyāt
bi-Tūnis

Malawi Human Rights
Commission

Malawi Human Rights
Resource Centre

MAP Foundation

Mardu iravunk'nerë
Hayastanum

Markaz al-Ma'lūmāt wa-al-
Ta'hīl li-Ḥuqūq al-Insān

al-Markaz al-Miṣrī li-Ḥuqūq
al-Mar'ah

Markaz al-Tawthīq wa-al-
l'ām wa-al-Takwīn fī Majāl
Ḥuqūq al-Insān

Browse websites by subject.

#

A-Z

Z-A



[Human rights \(General\)](#) 520

[Civil rights](#) 71

[Democracy](#) 42

[Women's rights](#) 41

[Ombudspersons](#) 30

[Civil society](#) 20

[Transitional justice](#) 19

[Indigenous peoples](#) 18

[Torture](#) 17

[Palestinian Arabs](#) 15

[Refugees](#) 15

[Rule of law](#) 15

[Truth commissions](#) 15

[Legal aid](#) 13

[Disappeared persons](#) 11

[Discrimination](#) 10

[Caste-based discrimination](#) 1

[Censorship](#) 1

[Chechens](#) 1

[Child abuse](#) 1

[Child labor](#) 1

[Child pornography --
Prevention](#) 1

[Child prostitution --
Prevention](#) 1

[Child sexual abuse --
Prevention](#) 1

[Child soldiers](#) 1

[Child trafficking](#) 1

[Child welfare](#) 1

[Children -- Legal status,
laws, etc](#) 1

[Children -- Services for](#) 1

[Land tenure](#) 1

[Land use, Rural -- Social
aspects](#) 1

[Law reform](#) 1

[Legal assistance to the poor](#) 1

[Lesbians](#) 1

[Mass media policy](#) 1

[Mayas](#) 1

[Medical care -- Law and
legislation](#) 1

[Medical scientists -- Civil
rights](#) 1

[Métis -- Education](#) 1

[Migrant labor](#) 1

[Military assistance,
American](#) 1

[Military courts](#) 1

HUMAN RIGHTS WEB ARCHIVE

CENTER FOR HUMAN RIGHTS DOCUMENTATION & RESEARCH AT COLUMBIA UNIVERSITY

Showing item 1 of 1.

Voix des sans-voix pour les droits de l'homme



Creator	Voix des sans-voix pour les droits de l'homme (Organization)
Organization Type	Non-governmental organizations
Organization Based In	Congo (Democratic Republic)
Geographic Focus	Congo (Democratic Republic)
Subjects	Human rights (General)
Summary	"La voix des sans voix est un ONG de défense des droits de l'homme en République démocratique du Congo."
Languages	French



[View site captures for http://www.vsv-rdc.org/](http://www.vsv-rdc.org/)

[View site captures for http://www.vsv-rdc.com/](http://www.vsv-rdc.com/)

[View site captures for http://www.congonline.com/vsv/](http://www.congonline.com/vsv/)

HRWA portal : Site details view, with links to crawl calendar page(s)

Displaying sites 1 - 30 of 614 (0.011 seconds)

« Previous 1 2 3 4 ... Next »

Relevance A-Z Z-A Results per page



5-18 Kinyom Chaedan

Site Details



ABONG--Associação Brasileira ...

Site Details



Abuelas de Plaza de Mayo

Site Details



Ação Brasileira pela Nutrição e ...

Site Details



Access to Justice

Site Details



Action Group for Health, Human...

Site Details

Narrow These Results

GEOGRAPHIC FOCUS

- [Global focus] (67)
- Israel (26)
- West Bank (23)
- Egypt (22)
- Gaza Strip (22)
- more »

LANGUAGE

- English (468)
- Spanish (110)
- Arabic (102)
- French (79)
- Russian (30)
- more »

ORGANIZATION BASED IN

- United States (56)
- United Kingdom (40)
- Egypt (27)
- Israel (15)
- India (14)
- more »

ORGANIZATION TYPE

- Non-governmental organizations (475)
- National human rights institutions (88)
- Other organization types (38)
- Individual site creators (12)

Default search is Site search (Metadata): blank search shows collection overview



HRWA portal : Search website descriptions (metadata)

Displaying **all 5 sites** (0.003 seconds) for: **trafficking** 



SANTAC

Southern Africa Regional Network against **Trafficking** and Abuse of Children...Southern Africa Regional

[Site Details](#) 



ECPAT International

ECPAT International is a global network of organisations and individuals working together for the

[Site Details](#) 



Hotline for Migrant Workers

"The Hotline for Migrant Workers (HMW), established in 1998, is a non-partisan, not for profit organization,

[Site Details](#) 



WOREC Nepal

"Women's Rehabilitation Centre (WOREC) works in partnership with grassroots people in order to resolve



Global slavery index

"The inaugural edition of the Global Slavery Index 2013 provides a ranking of 162 countries around the world,

Narrow These Results

GEOGRAPHIC FOCUS

[\[Global focus\]](#) (2)
[Africa, Southern](#) (1)
[Israel](#) (1)
[Nepal](#) (1)

LANGUAGE

[English](#) (5)
[French](#) (1)
[Hebrew](#) (1)
[Nepali](#) (1)
[Portuguese](#) (1)
[more](#) »

ORGANIZATION BASED IN

[Australia](#) (1)
[Israel](#) (1)
[Mozambique](#) (1)
[Nepal](#) (1)
[Thailand](#) (1)

ORGANIZATION TYPE

[Non-governmental organizations](#) (5)

SUBJECT

[Human trafficking](#) (4)
[Human rights \(General\)](#) (2)
[Child abuse](#) (1)
[Child labor](#) (1)

The image shows the top section of the HRWA portal. On the left, the text "HRWA" is displayed in white on a dark blue background. To its right is a search input field containing the word "trafficking". Further right is a "Pages" dropdown menu and a "Search" button with a magnifying glass icon. Below the search bar is a banner with a textured background. The banner contains the text "RIGHTS WEB ARCHIVE" in large, stylized letters. Below this, in smaller text, it says "STATION & RESEARCH AT COLUMBIA UNIVERSITY". On the right side of the banner, the word "BETA" is written in red. A search dropdown menu is open, showing two options: "Website descriptions" with an unselected radio button and "Page full text" with a selected radio button.

HRWA portal : Search page full text

Displaying page 1 for about **2,000,000** archived web page results (10.039 seconds) for: **trafficking** 

« Previous 1 2 3 4 ... Next »

Results per page ▾

[Untitled]

Sep 3, 2011 at 2:17AM UTC

<http://www.hrln.org/hrln/pdf/trafficking/VidabhaBhaskar.pdf>

Size: 11 MB - PDF

Domain- Domain+

Trafficking and the Law, Second edition

Jan 5, 2014 at 5:36AM UTC

<http://www.hrln.org/hrln/publications/books/928-trafficking-and-the-law-second-edition.pdf>

Trafficking and the Law, Second edition Trafficking and the Law, Second edition 1 / 1 ...

Size: 235 KB - PDF

Other Relevant Captures

- Mar 28, 2014 at 6:03PM UTC (235 KB)
- Jun 25, 2014 at 3:35AM UTC (235 KB)
- Oct 2, 2014 at 8:32PM UTC (235 KB)

Domain- Domain+

Trafficking and the Law, Second edition

Mar 28, 2014 at 10:01AM UTC

<http://hrln.org/hrln/publications/books/928-trafficking-and-the-law-second-edition.pdf>

Trafficking and the Law, Second edition Trafficking and the Law, Second edition 1 / 1 ...

Size: 235 KB - PDF

Other Relevant Captures

- Jun 24, 2014 at 6:57PM UTC (235 KB)
- Oct 1, 2014 at 2:19PM UTC (235 KB)
- Oct 1, 2014 at 5:04PM UTC (235 KB)

Narrow These Results

DOMAIN

[camp.org.pk](#) (397,310)
[suhakam.org.my](#) (146,840)
[humanrights.asia](#) (134,680)
[khpg.org](#) (134,574)
[errc.org](#) (121,384)
[more »](#)

DATE OF CAPTURE

[2012](#) (466,593)
[2011](#) (443,870)
[2010](#) (423,835)
[2014](#) (335,357)
[2013](#) (262,824)
[more »](#)
[custom range »](#)

FILE TYPE

[HTML](#) (1,881,269)
[PDF](#) (51,879)
[XML](#) (13,563)
[Document](#) (3,288)
[Presentation](#) (97)
[Spreadsheet](#) (19)

GEOGRAPHIC FOCUS

[\[Global focus\]](#) (466,753)
[Africa, Southern](#) (398,309)
[Asia](#) (164,522)
[Ukraine](#) (135,003)
[Africa](#) (119,565)
[more »](#)

DOMAIN

[amnesty.org.au](#) (118,067)
[amnesty.ca](#) (109,462)
[humanrights.asia](#) (85,877)
[iranhumanrights.org](#) (62,535)
[ishr.org](#) (62,373)
[more »](#)

DATE OF CAPTURE

[2011](#) (415,428)
[2012](#) (279,309)
[2010](#) (183,085)
[2008](#) (8,213)
[2009](#) (6,955)
[custom range »](#)

FILE TYPE

[HTML](#) (854,260)
[PDF](#) (36,585)
[Document](#) (1,405)
[XML](#) (701)
[Presentation](#) (27)

Displaying page 1 for about 890,000 archived web page results (5.266 seconds) for: **death penalty** 

« Previous 1 2 3 4 ... Next »

Results per page ▾

Death Penalty

Jun 2, 2011 at 9:47PM UTC

http://www.pchrgaza.org/portal/en/index.php?option=com_content&view=category&id=62:pchrdeathpenalty&Itemid=210

[Death Penalty...](#) [Death Penalty...](#)

Size: 40.5 KB - HTML

Other Relevant Captures

[Sep 2, 2011 at 10:40PM UTC \(40.2 KB\)](#)
[Mar 2, 2012 at 10:40PM UTC \(42.5 KB\)](#)
[Dec 2, 2011 at 10:37PM UTC \(41.4 KB\)](#)
[Apr 22, 2010 at 9:16PM UTC \(41.6 KB\)](#)
[Jun 2, 2010 at 7:36PM UTC \(41.1 KB\)](#)
[Dec 2, 2010 at 8:03PM UTC \(40.9 KB\)](#)
[Sep 2, 2010 at 8:47PM UTC \(41.6 KB\)](#)
[Jun 3, 2012 at 12:10AM UTC \(42.6 KB\)](#)
[Mar 2, 2011 at 8:40PM UTC \(40.5 KB\)](#)

[Domain-](#) [Domain+](#)

HRWA trial feature: search extension

You just searched in HRWA for: **death penalty**

Try your search in one of these related resources. (Search results will display in a new window.)

Archive-It

All web archives created by Archive-It partner institutions, including Columbia University.



HuriSearch

Custom live web search engine for over 5,000 human rights websites powered by Huridocs.



ArchiveGrid

Database of nearly 2,000,000 archival collection descriptions.



Office of the High Commissioner for Human Rights, United Nations

Documents, news, and publications on the OHCHR website.



Universal Human Rights Index, maintained by the OHCHR

Information issued by the UN's international human rights mechanisms: Treaty Bodies, Special Procedures, and the Universal Periodic Review.



HRWA trial feature : Search extension

Displaying page 1 for about 2,200 archived web page results (0.013 seconds) for: **chippewa** 

SEARCH EXPANSION TRIAL FEATURE

Expansion Off Expansion On

Expanded search terms were found for this query!

« Previous 1 2 3 4 ... Next »

Results per page ▾

The Chippewas of Saugeen

Jun 2, 2010 at 8:51PM UTC

<http://www.peacebrigades.org/archive/nap/nap96-05.html>

The **Chippewas** of Saugeen... The **Chippewas** of Saugeen | HomePage | The **Chippewas** of Saugeen Peace Brigades International Special Report: December, 1996 Two Peace Brigades International (PBI) volunteers recently acted as observers during a meeting involving the **Chippewas** of Saugeen. The Saugeen are the sister band of the **Chippewas** of Nawash, with whom PBI has been involved
Size: 4.3 KB - HTML

Other Relevant Captures

Sep 2, 2010 at 9:54PM UTC (4.26 KB)

See all results from [peacebrigades.org](#)

Domain- Domain+

Observation Teams for the Chippewas of Nawash

Jun 2, 2010 at 8:51PM UTC

<http://www.peacebrigades.org/archive/nap/nap96-04.html>

Observation Teams for the **Chippewas** of Nawash... Observation Teams for the **Chippewas** of Nawash | HomePage | Observation Teams for the **Chippewas** of Nawash Peace Brigades International Special Report: September, 1996 At the end of August, Peace Brigades

Narrow These Results

DOMAIN

[docip.org](#) (1,398)
[peacebrigades.org](#) (318)
[humanrights.gov.au](#) (86)
[we-change.org](#) (84)
[iwgia.org](#) (76)
[more »](#)

DATE OF CAPTURE

2010 (1,602)
2011 (430)
2012 (112)
2008 (9)
2009 (1)
[custom range »](#)

FILE TYPE

HTML (1,943)
PDF (207)
Document (4)

GEOGRAPHIC FOCUS

[\[Global focus\]](#) (1,865)
[Australia](#) (86)

HRWA trial feature : Search expansion (include synonyms of names of peoples)

Displaying page 1 for about 4,500 archived web page results (0.193 seconds) for: **chippewa**

SEARCH EXPANSION TRIAL FEATURE

Expansion Off Expansion On

Your search has been expanded to include the following terms:

• Algic, Anishinabe, Bawichtigoutek, Bungee, Bungi, Chipouais, Lac Courte Oreilles, Ochepwa, Odjibway, Ojebwa, Ojibua, Ojibwa, Ojibwauk, Ojibway, Ojibwe, Otchilpwe, Salteaux, Saulteaux (via "**chippewa**")

Equivalent Query: (chippewa OR Algic OR Anishinabe OR Bawichtigoutek OR Bungee OR Bungi OR Chipouais OR "Lac Courte Oreilles" OR Ochepwa OR Odjibway OR Ojebwa OR Ojibua OR Ojibwa OR Ojibwauk OR Ojibway OR Ojibwe OR Otchilpwe OR Salteaux OR Saulteaux)

« Previous 1 2 3 4 Next »

Results per page ▾

multi-ojibway

May 10, 2010 at 5:46PM UTC

<http://www.chrc-ccdp.ca/pdf/multi-ojibway.pdf>

multi-[ojibway](#)... multi-[ojibway](#) [Ojibway](#) / [Ojibway](#) For more information, contact the regional office of the Canadian Human Rights Commission nearest you: Atlantic/Atlantique Halifax : (902) 426-8380 Toll Free/Sans frais : 1-800-999-6899 TTY/ATS : 1-888-643-3304 Fax/Télocopieur : (902) 426-2685 Quebec/Québec Montreal/Montréal : (514) 283-5218 Toll Free/Sans frais : 1-800-999-6899...
Size: 138 KB - PDF

Other Relevant Captures[Sep 2, 2010 at 7:33PM UTC \(138 KB\)](#)**Narrow These Results**

DOMAIN

[docip.org](#) (2,711)
[chrc-ccdp.ca](#) (458)
[peacebrigades.org](#) (318)
[amnesty.ca](#) (139)
[we-change.org](#) (122)
[more »](#)

DATE OF CAPTURE

[2010](#) (2,935)
[2011](#) (1,156)
[2012](#) (329)
[2008](#) (21)
[2009](#) (14)
[custom range »](#)

FILE TYPE

[HTML](#) (4,092)
[PDF](#) (359)
[Document](#) (4)

GEOGRAPHIC FOCUS

[\[Global focus\]](#) (3,340)

HRWA trial feature : Search expansion (include synonyms of names of peoples)

HRWA Search / Solr details

HRWA running on Solr 4.2

(not ready to update to Solr 5.0, but 4.10 is a reasonable goal)

HRWA Solr Configuration

HRWA is a Blacklight site--Solr configuration is based on the default blacklight configuration: <https://github.com/projectblacklight/blacklight/wiki/Solr-Configuration>

(HRWA is running on Blacklight 3, and the latest Blacklight version is 5.12.)

Using explicitly named fields rather than dynamic fields (<https://cwiki.apache.org/confluence/display/solr/Dynamic+Fields>), but will likely change to dynamic fields during the next application update

We allocate 16GB of RAM to the Tomcat server that hosts our Solr 4.2 Webapp.

Two Separate Cores

HRWA has two types of search:

Site Search: Searches through 600+ Voyager records, each representing a crawled site. Facets and displayed metadata come from Voyager MARC.

Fulltext Search: Searches through over 50 Million fulltext documents (HTML, PDF, Document, XML, Spreadsheet, Presentation), with added facets from merged-in Voyager record data.

In order to keep our Site Search speed from slowing down as we add more fulltext documents, we have chosen to use two solr cores rather than one. The **fulltext core** can grow indefinitely, but **site core** search speed will be unaffected.

Search Parameters

The Human Rights Web Archive application can be found here in a public repository on GitHub: https://github.com/cul/hrwa_blacklight

Our current Site Search parameters can be seen here (starting on line 9):

https://github.com/cul/hrwa_blacklight/blob/master/lib/hrwa/find_site_search_configurator.rb#L9

(The configuration is pretty basic.)

Our current Fulltext search parameters can be seen here (starting on line

9): https://github.com/cul/hrwa_blacklight/blob/master/lib/hrwa/archive_search_configurator.rb#L9

Known Factors that Affect Performance

Result Grouping

Grouping multiple site capture results by common URL is a useful feature, but comes at a performance cost.

Searches take about 4x longer because of grouping. We may want to consider disabling grouping in the future and finding a different way to show alternate site captures. Redundantly storing URLs for all captures in each search result may be a good option.

More information about result grouping: <https://cwiki.apache.org/confluence/display/solr/Result+Grouping>

Search Term Highlighting

When users perform fulltext searches, we offer text snippets with term highlighting. This is useful, but also leads to a performance hit.

<https://cwiki.apache.org/confluence/display/solr/Highlighting>

Opportunities for Improvement (summary + additional info)

- Update to the latest version of Java (Java 8). Our prior work was based on Java 6, which is old and no longer supported.
- Use a NoSQL database like MongoDB rather than MySQL, which could be faster overall, and could offer easier opportunities for horizontal scalability through a distributed setup.
- Invest in Solid State drives. Solr and our databases would be faster.
- Update to the latest version of Solr and look into using SolrCloud for a distributed setup. Further leverage horizontal scalability.
- Don't use Solr grouping because it is slow. Redundantly store grouped data about alternate crawls in each record. This would lead to more work on the indexing side, but faster results on the searching side.
- Maybe start hashing the resource content for individual page crawls and compare hashes when indexing to avoid duplicate records?

AWG full-text search session questions

... beyond just using Solr, are there configuration and schema options, relevancy weightings, results groupings, indexing architectures, front-end interfaces, software packages, etc. that we are or could be using in common?

Some details about our implementation:

- Blacklight 3 [front-end app]
- Solr 4.2 [backing data source]
- MySQL [for intermediate storage of unpacked WARC data]
- custom Java app + Heritrix Java library [for extracting (W)ARCs and indexing data].

AWG full-text search session questions

... What do we know about (web archive) users' needs and expectations for full-text search?

Not that much!

But we are currently conducting some in-depth user interviews with human rights researchers to help learn more about this.

AWG full-text search session questions

What could we work on together that would make Solr-based full-text search for web archives easier to implement?

Community-built documentation about workflows and "good fit" open source projects for handling various stages of archive file indexing and searching. For example:

- Solr vs. ElasticSearch for indexing/searching
- Recommended databases for intermediate storage of WARC file data before indexing into Solr (Relational vs. NoSQL – e.g. MySQL vs. MongoDB)
- Community-managed lists of the most useful/common file extensions / mime types, and which extensions should be used/ignored for different scenarios.
- Community-managed wiki for best indexing practices / strategies. e.g. Special rules for indexing HTML docs? Special rules for detecting redundant pages / content.

AWG full-text search session questions

... What subset of challenges for web archives and Solr are shared with the larger community of Solr users?

Indexing Performance - We've been limited by IO more than anything. Our indexer is a multithreaded Java application, but we can't take advantage of all of the cores on the machine because of IO limits. More memory would also be useful so that we could hit the disk less often when doing concurrent processing of large files.

Search Performance - Too many documents means long searches. Would like to tune Solr (and eventually set up distributed SolrCloud) for better performance.

Thanks!

Questions/comments about HRWA collection development, permissions policy, descriptive metadata, portal functionality

Contact Alex: at2186@columbia.edu

Questions/comments about HRWA Solr/Blacklight configuration, other technical info

Contact Eric: elo2112@columbia.edu

