Web Archiving in Context and Up Close

May 06, 2011
Digital Library Seminar
Columbia University Libraries
Mellon Project on Web Resource Collection Program Development

Presentation Overview

- How it works...
  1. Crawler: what it is, how it works
  2. Crawler + website, how they interact
  3. Challenges
  4. Products
  5. Reports
  6. Rendering
- Website walk-through
  1. Successfully crawled site/live site
  2. Site with navigation problems
  3. Site with search box
  4. Site with flash
  5. Blogs/linked content
  6. Social media
  7. News aggregators
- Tools/services
  - Subscription
  - Corporate/E-Discovery/Compliance
- Other web archiving tools

Tools, Resources, Process

Tools & Resources

- Crawler
- Storage for crawler-generated files
- Software to render files
- Access via portal, catalog, index, etc.

Process

- Select
- Test
- Scope
- Crawl
- Review
- Provide Access

Crawlers

Example: Heritrix

What does a crawler do?
Starting from a particular URL, a crawler will gather information about that starting page, while looking for additional URLs. When it finds a new URL, it will jump to each to that URL and repeat the processing of gathering information and looking for additional URLs.

What users see on the live site…

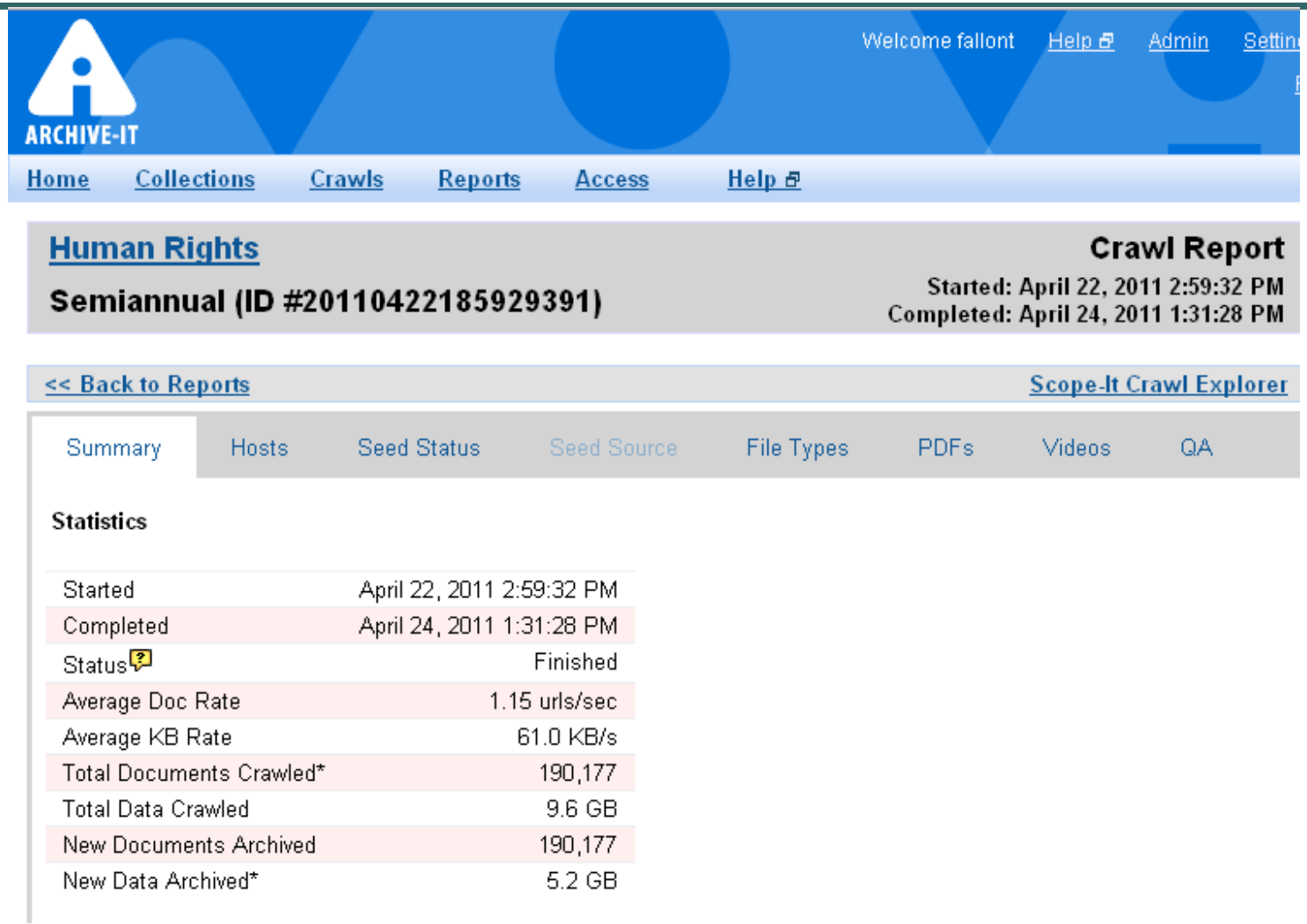
What a crawler sees…

```
dns:amnestymauritius.org
http://amnestymauritius.org/
http://amnestymauritius.org/downloads/didi1.mp3
http://amnestymauritius.org/downloads/didi2.mp3
http://amnestymauritius.org/downloads/didienglish.doc
http://amnestymauritius.org/favicon.ico
http://amnestymauritius.org/french/1.2.6
http://amnestymauritius.org/french/Microsoft.XMLHTTP
http://amnestymauritius.org/french/application/x-www-form-urlencoded
http://amnestymauritius.org/french/become.php
http://amnestymauritius.org/french/becomemada.php
http://amnestymauritius.org/french/contact.php
http://amnestymauritius.org/french/downloads.php?cat_id=1
http://amnestymauritius.org/french/downloads.php?cat_id=2
http://amnestymauritius.org/french/downloads.php?cat_id=3
http://amnestymauritius.org/french/downloads.php?cat_id=4
http://amnestymauritius.org/french/downloads.php?cat_id=5
http://amnestymauritius.org/french/downloads.php?cat_id=6
http://amnestymauritius.org/french/downloads.php?page_id=30
http://amnestymauritius.org/french/downloads/CONSTITUTION_OF_AIMS_AS_ADOPTED_BY_EGM_Feb09KP_RP.pdf
http://amnestymauritius.org/french/downloads/Statute_of_Amnesty_International.pdf
http://amnestymauritius.org/french/downloads/manifeste2010.doc
http://amnestymauritius.org/french/downloads/maurice_constitution.pdf
http://amnestymauritius.org/french/downloads/passport.pdf
http://amnestymauritius.org/french/downloads/statuteAmnesty.doc
http://amnestymauritius.org/french/images/01.jpg
http://amnestymauritius.org/french/images/02.jpg
http://amnestymauritius.org/french/images/03.jpg
http://amnestymauritius.org/french/images/04.jpg
http://amnestymauritius.org/french/images/05.jpg
http://amnestymauritius.org/french/images/06.jpg
http://amnestymauritius.org/french/images/07.jpg
http://amnestymauritius.org/french/images/08.jpg
```

WARC Files & Crawl Reports

- WARC file:
  - Web ARChive format. "*Used for web-accessible content in archived state, representing the final form disseminated in final state over the web to a user agent (web browser).*" (LC Digital Preservation)
- Archive-It Report
  - Summary
  - Hosts
  - Seed status
  - Seed source
  - File types
  - PDFs
  - Videos

Archive-It Crawl Report



Hosts Report (Archive-It)

View only hosts containing [ ] Filter Clear « First « Previous Next » Last » 1 through 100 of 2493

Host	URLs	Data	New URLs	New Data	Queued	Robots.txt Blocked	Out of Scope
amnesty.or.jp	95,661	2.4 GB	7,755	198.3 MB	0	0	781
www.amnesty.ca	25,781	1.8 GB	13,359	604.6 MB	0	0	0
www.amnesty.ch	24,987	1.1 GB	12,993	561.8 MB	0	0	5
cdhrap.net	19,026	1.3 GB	17,149	1.1 GB	0	0	22,800
woeser.middle-way.net	9,850	1.4 GB	9,850	1.4 GB	0	81	1,033
www.blogger.com	3,104	39.8 MB	3,097	39.7 MB	0	0	26,263
amnesty.ca	2,472	276.0 MB	945	19.3 MB	0	0	0
2.bp.blogspot.com	1,447	27.6 MB	1,446	27.5 MB	0	0	1,146
3.bp.blogspot.com	1,437	27.7 MB	1,436	27.7 MB	0	0	1,122
4.bp.blogspot.com	1,401	27.8 MB	1,400	27.8 MB	0	0	1,110
1.bp.blogspot.com	1,383	27.1 MB	1,382	27.1 MB	0	0	1,096
www.andalusitas.net	972	28.1 MB	972	28.1 MB	0	0	0
www.ombudsman.org.na	588	43.6 MB	588	43.6 MB	0	0	0
tamaynut.org	346	5.1 MB	346	5.1 MB	0	0	0

Wayback Machine (Internet Archive)

"Alexa Internet, in cooperation with the Internet Archive, has designed a three dimensional index that allows browsing of web documents over multiple time periods, and turned this unique feature into the Wayback Machine." (Internet Archive)

COLUMBIA UNIVERSITY LIBRARIES

Human Rights Web Archive (Columbia University Libraries)

INTERNET ARCHIVE WayBackMachine

Enter Web Address: http:// All Take Me Back Compare Archive Pages

Searched for http://amigosdemujeres.org/ 6 Results RSS Metadata

Look up URL in general Internet Archive web collection Proxy Mode Help

* denotes when page was updated

Search Results for Jan 1, 2005 - Dec 31, 2011						
2005	2006	2007	2008	2009	2010	2011
0 pages	0 pages	0 pages	0 pages	1 page	4 pages	1 page
				May 14, 2009 *	Mar 19, 2010 * Jun 2, 2010 * Sep 2, 2010 * Dec 2, 2010 *	Mar 2, 2011 *

Anatomy of a Website

- Public-facing content (what a user is meant to see)
- Underlying content
  - Text
  - Images
  - Databases
  - Scripts
  - External software
- Text/html
- Documents: PDFs, .ppt, .xls, .doc,
- Audio/visual files
  - jpeg, gif, png, mpeg, avi, x-flv, mp4, x-ms-wmx

Challenges

- Dynamic pages
- Flash
- Javascript
- Characters/fonts
  - ex: URLs with Korean characters
- Databases
- News aggregators
- Robots.txt
- Crawler traps
- Social media

Successfully crawled site…

Home | About | Contact us | Sitemap | عربي | French | RSS Feeds

Search »go | Advanced Search

Civil Society Petition: General Assembly should suspend Libya's UN Human Rights Council membership...

- Home
- HR Advocacy
  - HR in the Arab World
  - International & Regional Mechanisms
- HR Education & Dissemination
  - Publications
  - Sawasia Magazine
  - Ibn Rushd Salon
  - Training & Educaiton

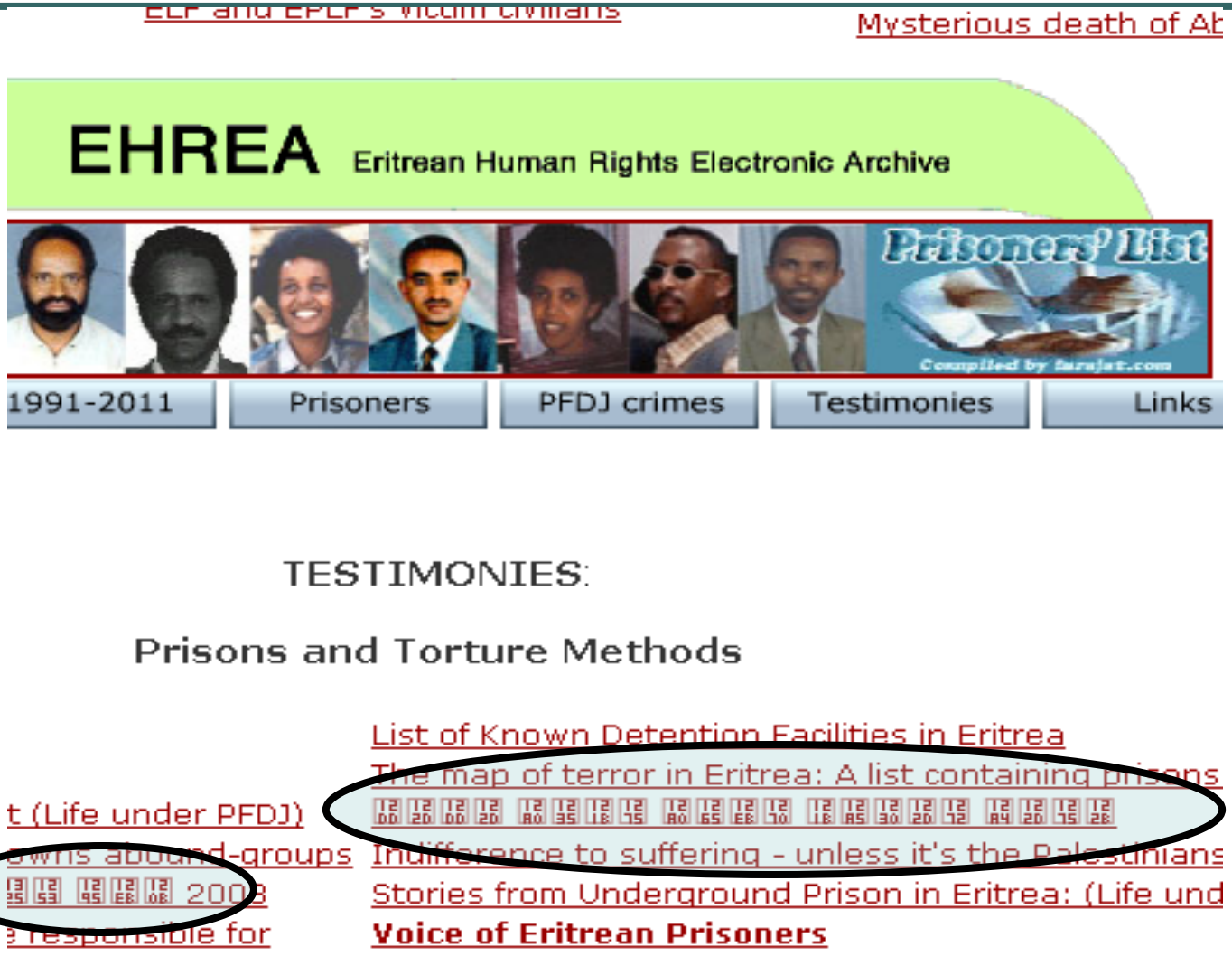
Publications

Towards a Democratic Legislation Supporting the Independence of Non Governmental Organizations (NGOs) | 15/03/2009

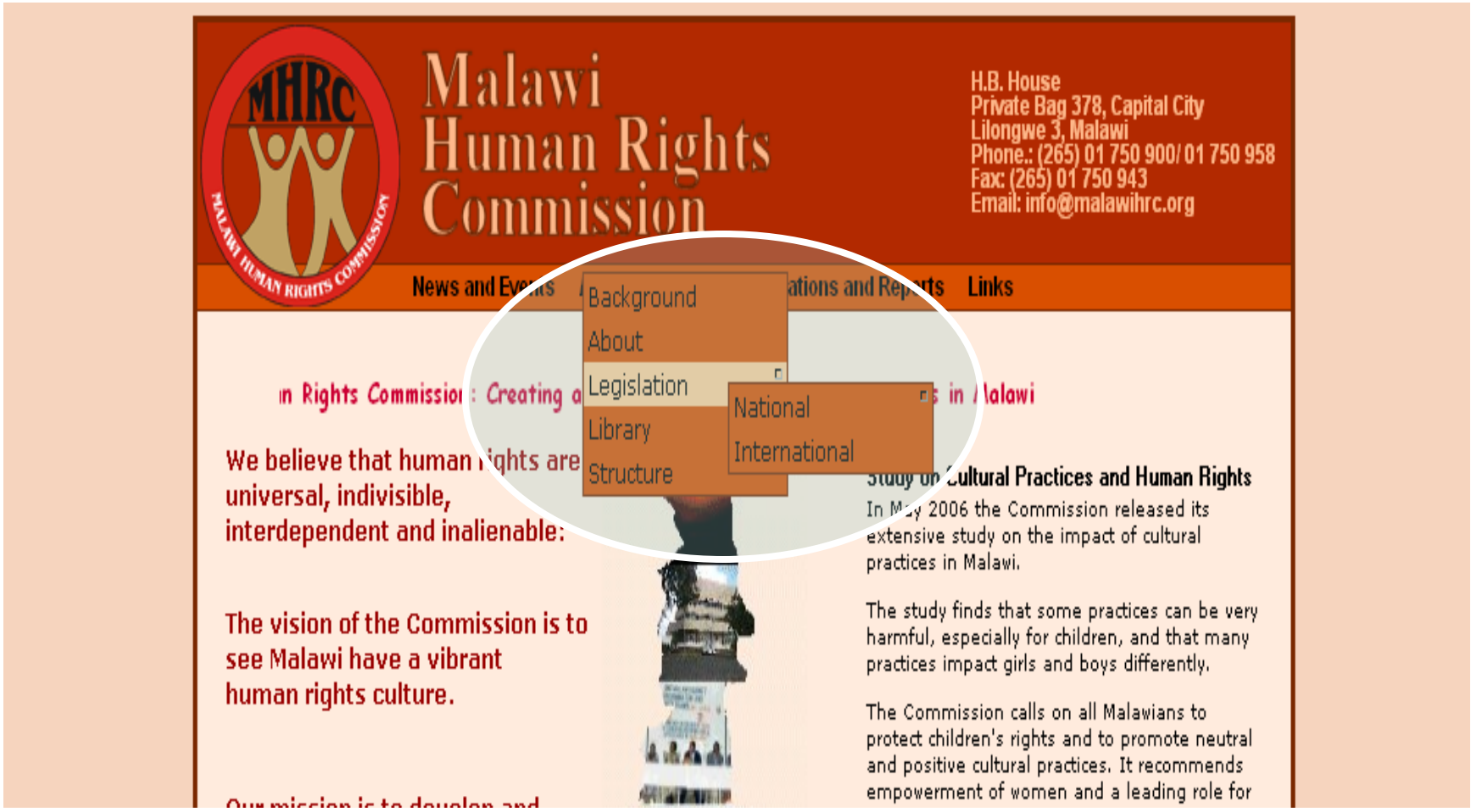
Editor: Essam El Din M. Hassan

The significance of the study becomes clear in light of the current moves by the government to amend the present law, something that will probably result in imposing additional restrictions on NGOs. In particular, the official moves to amend the law come at a time that is witnessing extensive legislative attacks on liberties, and increasing tendency to oppress the rights of expression and association and other forms of political and social mobility. It is worth noting that the deliberations regarding amending the law came simultaneously as the closing down of two human rights organizations for the first time since the establishment of human rights organizations.

Characters/fonts



Navigation


Malawi Human Rights Commission
H.B. House
Private Bag 378, Capital City
Lilongwe 3, Malawi
Phone.: (265) 01 750 900/ 01 750 958
Fax: (265) 01 750 943
Email: info@malawihrc.org
News and Events
Links
Background
About
Legislation
Library
Structure
National
International
We believe that human rights are universal, indivisible, interdependent and inalienable:
The vision of the Commission is to see Malawi have a vibrant human rights culture.
Study on Cultural Practices and Human Rights
In May 2006 the Commission released its extensive study on the impact of cultural practices in Malawi.
The study finds that some practices can be very harmful, especially for children, and that many practices impact girls and boys differently.
The Commission calls on all Malawians to protect children's rights and to promote neutral and positive cultural practices. It recommends empowerment of women and a leading role for

Searching/Databases

Flash


اللغة العربية | English
Updated: Tuesday December 28, 2010
The Palestinian Human Rights Monitoring Group
المجموعة الفلسطينية لمراقبة حقوق الانسان
Tax Exemption in the U.S.A
HomePage
Profile
Articles
The Monitor
Subscriptions
Volunteers
Contact us
tails
Related Links
Press Release
Past Projects
New Articles
The Role of Islam in the Israeli-Palestinian Conflict

Flash


Firefox prevented this site from opening a pop-up window.
Options
통합검색
SEARCH
HOME | INTRANET | SITEMAP
시각장애인 5·18학습사이트 ENGLISH 아카이브
5·18기념재단
재단소개
재단사업
5·18민중항쟁
알림마당
참여마당
열린마당
웹진마당
컨텐츠마당

Blogs


Delito Mayor
Desde aquí
El auditorio imbécil
El Blog de Dimas
El blogueo por el blogueo
El blogueo por el blogueo
El hombre en las nubes
El pequeño hermano
Habanemia
Indocubanos
Injusticia notoria
Kuba Sepia
La colmena
La rosa descalza
La voz del Morro
La voz del Morro
Los hijos que nadie quiso
Lunes de post Revolución
Mala letra
Mermelada
Mi isla al mediodía
Octavo Cerco
Pedimos la palabra
Re-Evolución
Reportes de viaje
Reportes de viaje
Retazos
Revista Voces
Esa mañana, la máquina de regadío estaba parada en medio del terreno y parecía un albatros de amplias alas atascado bajo el sol. Mis amigas y yo nos metimos en la cabina vacía, tocamos la palanca, los botones, el timón. Saltamos sobre el remendado asiento y fantaseamos con que aquel trozo de metal chirriante iba a echar andar y mojaríamos con su riego a todos los estudiantes. Nos reíamos por anticipado, pero ni una sola gota salió de los larguísimos tubos que se extendían a ambos lados. Sin embargo, mientras husmeábamos aquí y allá nos topamos una lata con unas frutas raras. Tenían la forma de un pimiento, pero el color iba del amarillo al anaranjado intenso y una semilla les colgaba por fuera. Jóvenes urbanas, atrapadas entre las carencias del racionamiento y el colapso agrícola, no había forma de que supiéramos que aquello era un "marañón".
Les hincamos el diente de inmediato. Dulce, suave y después, cuando la boca comenzó a resecársenos, pensamos que nos habíamos envenenado. Corrimos horrorizadas, gritando. La carcajada del profesor duró largos minutos. Cuando la sensación astringente pasó, nos quedó el deseo de morder otra vez esa textura que ya había sido cantada en las décimas guajiras, mencionada por nuestros abuelos y pintada por algunos pinceles del siglo anterior. Quedé impresionada por aquella fruta prohibida de nuestro paraíso socialista. Pasarían casi veinte años antes que la volviera a encontrar.
Share
Tweet
share
Mayo 3rd, 2011 | 283 comentarios | Imprimir
La casona, el país
de mi ultima negativa de viaje
posted about 9 hours ago
#cuba Oscar Elias Biscet en su casa
posted 7 days ago
#cuba#GY 7 numeros de la revista Voces
posted 24 days ago
#cuba#GY Presentacion Voces 7
posted 25 days ago
#GY#Cuba Presentacion de la revista Voces
posted 25 days ago
twitpic
Share Your Photos

Robots.txt

http://www.bgcentar.org.rs/robots.txt

```
User-agent: *
Disallow: /administrator/
Disallow: /cache/
Disallow: /components/
Disallow: /images/
Disallow: /includes/
Disallow: /installation/
Disallow: /language/
Disallow: /libraries/
Disallow: /media/
Disallow: /modules/
Disallow: /plugins/
Disallow: /templates/
Disallow: /tmp/
Disallow: /xmlrpc/
```

Social media…

Services: Subscription and OS/Free

- Subscription/Fee-based
  - Archive-It
  - CDL-Web Archiving Service
  - Archivethe.net
  - OCLC Web Harvester
- OS/Free; no support
  - Web Curator Tool
  - Netarchive Suite
  - Heritrix + Wayback

Corporate/E-Discovery/Compliance

- Iterasi
- Hanzo
- Reed Technology (affil. LexisNexis)

Other tools

- Scrapbook Plus
  - Add-on
  - Annotate, select specific content
- Diigo
  - Bookmarks, screenshots and annotation
- Read it Later
  - Offline copy of site
  - Add-on
- Zotero
  - Snapshot, links to live page
  - Locate in Wayback Machine function
- Bundlr, Delicious
  - Bookmark collections, link aggregators
- HTTrack
  - Offline copy of site
- GNU Wget (commandline tool)
  - Local version of site