

---

Billions and billions of  
objects, METS, PREMIS, oh  
my!

---

Gina Jones

Office of Strategic Information

Web Archiving Team

The Library of Congress >> More Online Collections

# Library of Congress Web Archives *Minerva*

BROWSE | SEARCH | TECHNICAL INFORMATION

[home](#)

## Web Archives Available:

- [Crisis in Darfur, Sudan, Web Archive, 2006](#)
- [Iraq War, 2003 Web Archive](#)
- [Law Library Legal Blawgs Web Archive](#)
- [Library of Congress Manuscript Division Archive of Organizational Web Sites](#)
- [Papal Transition 2005 Web Archive](#)
- [September 11, 2001 Web Archive](#)
- [United States 107th Congress Web Archive](#)
- [United States 108th Congress Web Archive](#)
- [United States Election 2000 Web Archive](#)
- [United States Election 2002 Web Archive](#)
- [United States Election 2004 Web Archive](#)
- [Visual Image Web Sites Archive](#)



The Library of Congress Web Archives (LCWA) is composed of collections of archived web sites selected by subject specialists to represent web-based information on a designated topic. It is part of a continuing effort by the Library to evaluate, select, collect, catalog, provide access to, and preserve digital materials for future generations of researchers. The early development project for Web archives was called MINERVA.

[home](#)

Done



---

# Tools and Formats

- Heritrix produces the content
    - Open source crawler developed by the Internet Archive in 2004
    - Outputs in WARC format
      - ISO 28500:2009
      - Early output-ARCs
  - Wayback “Viewer” provides the access
    - Open source tool developed by the Internet Archive
-

---

# Metadata Object Descriptive Schema

- Nomination database
    - URL
    - Access Rights
  - Metadata Extraction
    - Capture range
    - Document title
    - Page/site metadata
    - Kinds of documents
  - Catalogers provide LC subject headings
-

# Crisis in Darfur, Sudan Web Archive, 2006

<< [Back](#) [ [Display Archived Web Site](#) ]

**Title:** Africa Action: Activism for Africa Since 1953

**Date Captured:** February 20, 2006 - November 27, 2006  
[Archived Site](#)

**Subject(s):** Sudan--History--Darfur Conflict, 2003-Genocide

**Language(s):** English

**Genre:** Web site

**Access Condition:** None

**URL at time of capture:** [www.africaaction.org/index.php](http://www.africaaction.org/index.php)

**Citation ID:** <http://hdl.loc.gov/loc.natlib/mrva0011.0035>

**Record ID:** mrva0011.0035

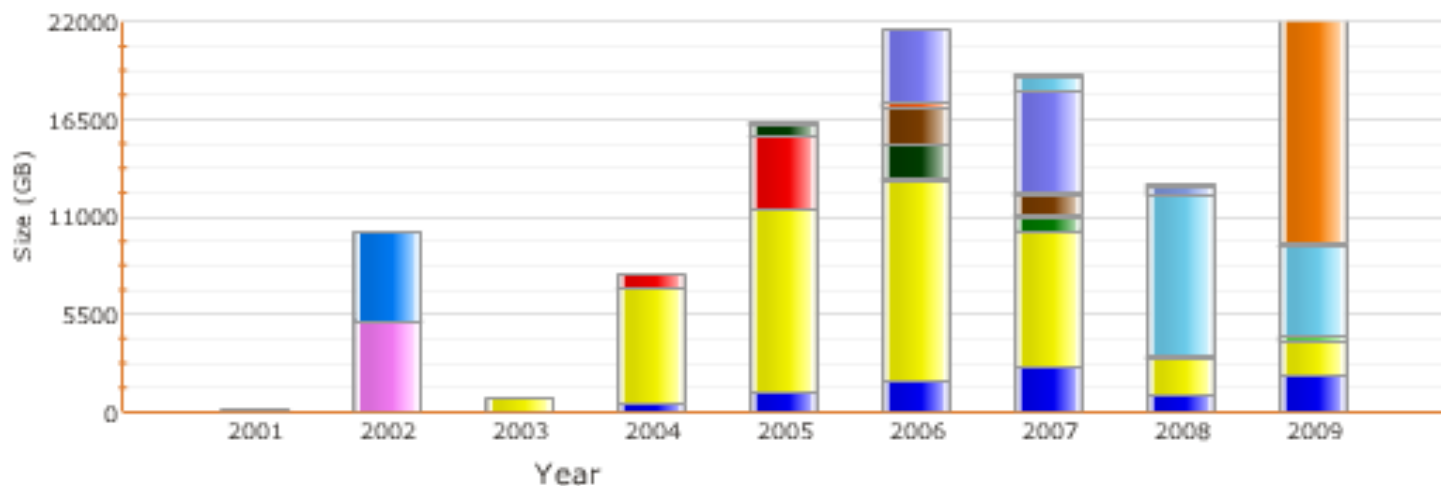
**Collection Title:** [Crisis in Darfur, Sudan Web Archive, 2006](#)

[Display MODS XML record](#)

[home](#) >> [overview](#) >> [browse results](#) >> **[bibliographic information](#)**

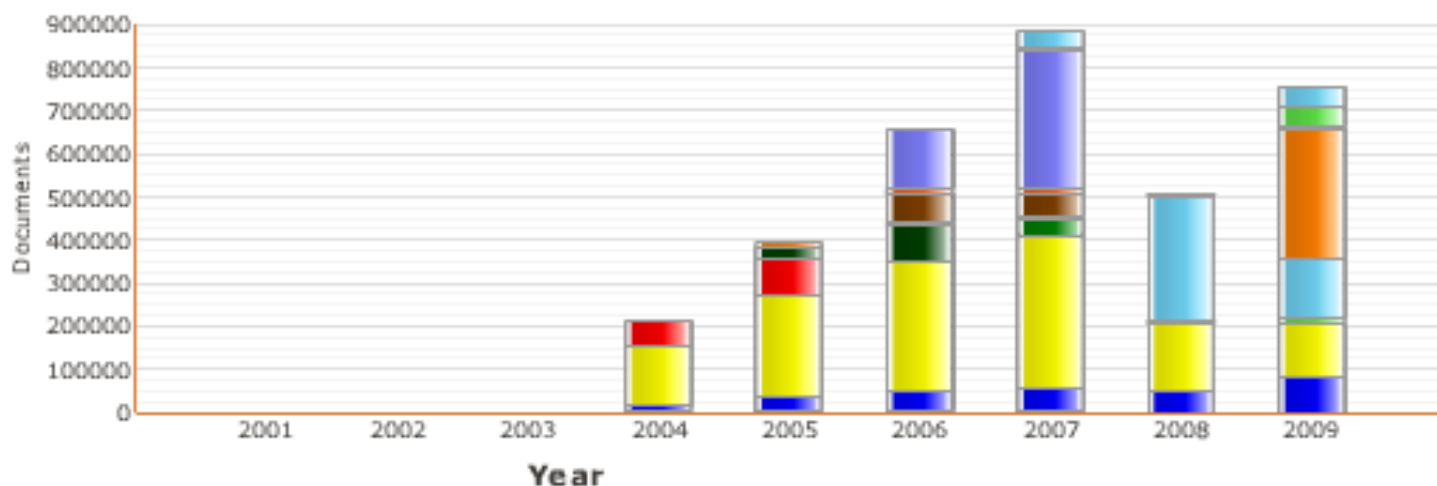
## Total Compressed Arc Size of Web Content

Total GB to Date = 121551



	2001	2002	2003	2004	2005	2006	2007	2008	2009	Totals
Election 2000	136									136
Election 2002		5000								5000
Olympics 2002		165								165
Sept. 11 & Remembrance		5056								5056
Congress			50	529	1142	1793	2561	881	2133	9088
Iraq War			712	6412	10272	11212	7667	2069	1900	40243
CRS Content				1						1
SMCCW Project				23						23
Election 2004				806	4113					4919
Univ. of AZ Terrorism					64	119	801			984
Supreme Court					541	1919				2460
Pope					232					232
Singlesite						33	22			55
Darfur						2091	1132			3223
Visual						19	7			25
Manuscript						256	188			443
Election 2006						4116	5664			9780
Legal Blogs							48	193	250	491
Election 2008							801	9075	5022	14898
Public Policy							113	470	128	711
Egypt 2008								175		175
Presidential Transition									21850	21850
Indonesia Gen Elec 2009									578	578
Overseas Operations									1015	1015
<b>Totals</b>	<b>136</b>	<b>10221</b>	<b>762</b>	<b>7771</b>	<b>16364</b>	<b>21557</b>	<b>19003</b>	<b>12863</b>	<b>32875</b>	<b>121551</b>

## Total Documents Crawled In Thousands



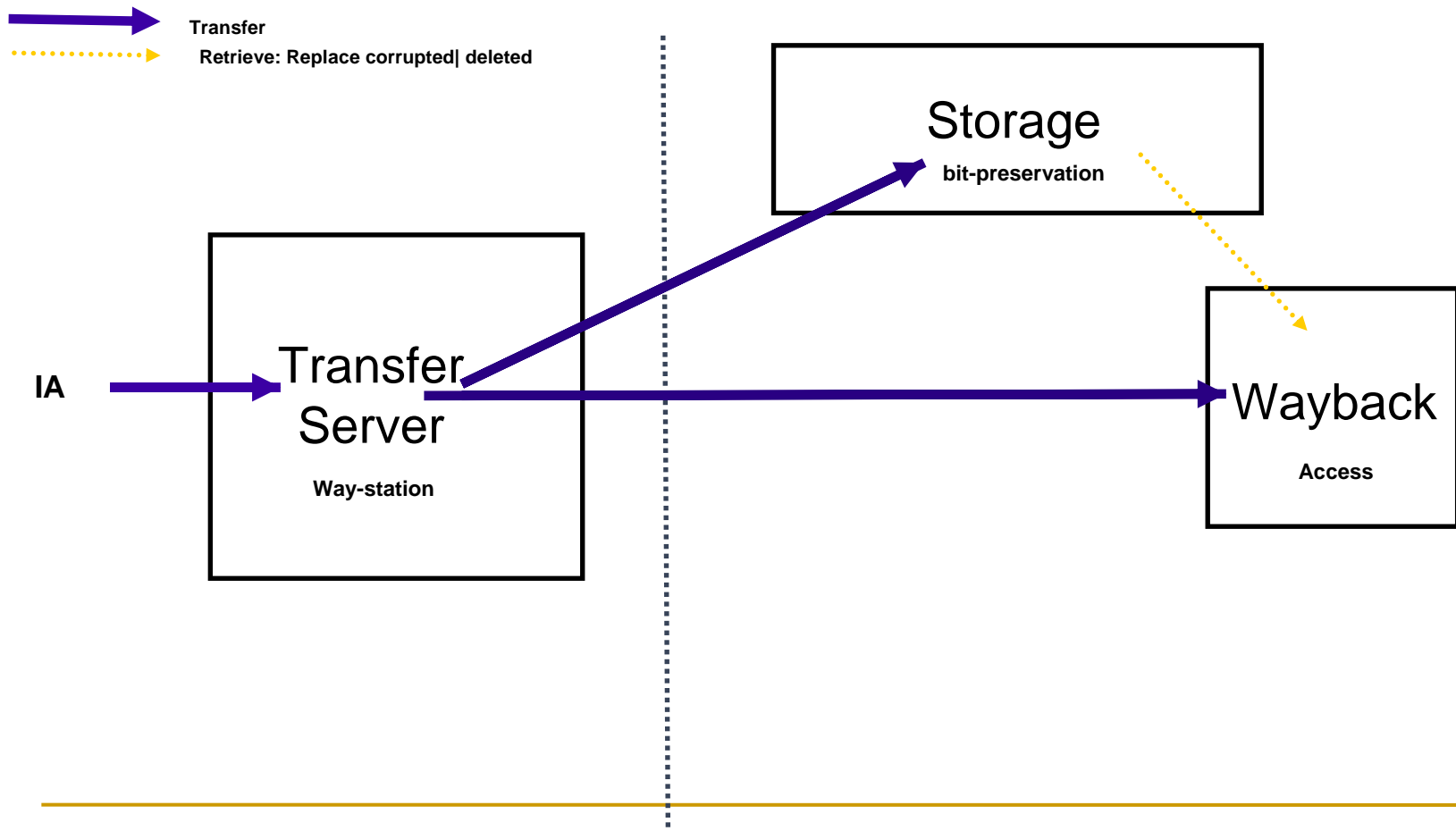
	2001	2002	2003	2004	2005	2006	2007	2008	2009	Totals
Election 2000	0									0
Election 2002		0								0
Olympics 2002		0								0
Sept. 11 & Remembrance		0								0
Congress			0	14108785	33425876	50373384	57213135	51188589	81210074	287519843
Iraq War			0	139961964	237627563	296915856	353531329	154195619	122813721	1305046052
CRS Content				0						0
SMCCW Project				0						0
Election 2004				56891439	81785070					138676509
Univ. of AZ Terrorism					0	0	37853659			37853659
Supreme Court					27875086	87105426				114980512
Pope					11325596					11325596
Singlesite						2908614	2427333			5335947
Darfur						66974552	0			66974552
Visual						319249	93781			413030
Manuscript						13374953	13267100			26642053
Election 2006						135813310	322628447			458441757
Legal Blogs							6475715	10302927	14200411	30979053
Election 2008							37853659	286374377	140151996	464380032
Public Policy							806471	2642973	450216	3899660
Egypt 2008								5296451		5296451
Presidential Transition									300033466	300033466
Indonesia Gen Elec 2009									50276613	50276613
Overseas Operations									45012063	45012063
Totals	0	0	0	210962188	392039191	653785344	832150629	510000936	754148560	3353086848

---

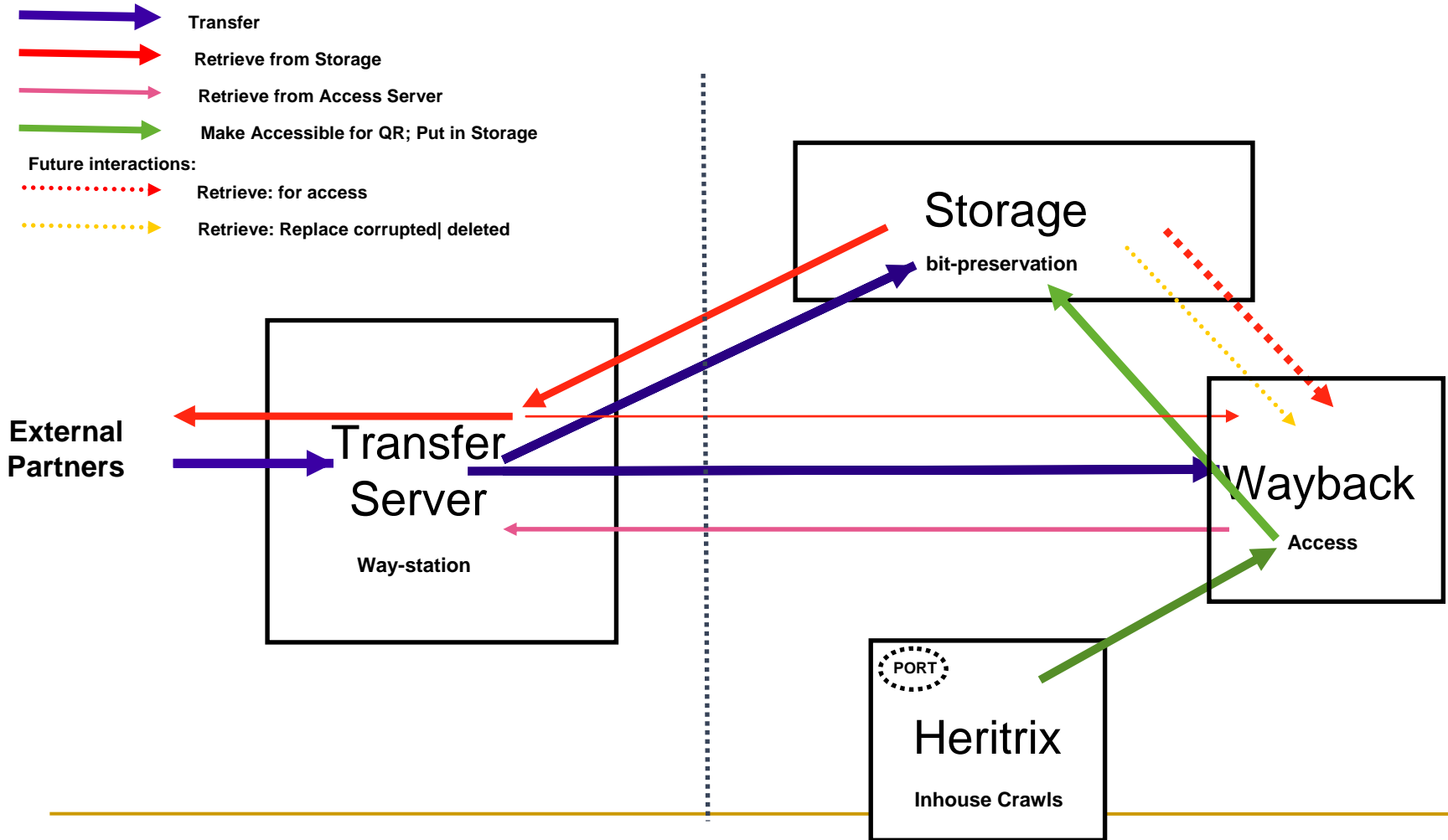
# Content-A “Plain Vanilla” Use Case

- **File & Content is fixed**
    - What Heritrix writes = what we transfer = what we store = what we make accessible to the public
    - No masters or derivatives
  - **Organization**
    - Just files in buckets
  - **Bagit file Packaging Format**
    - Bags persist from transfer to storage to access
-

# 2008: Interactions & Workflows



# ...and here we are in 2009



---

# Web Archives

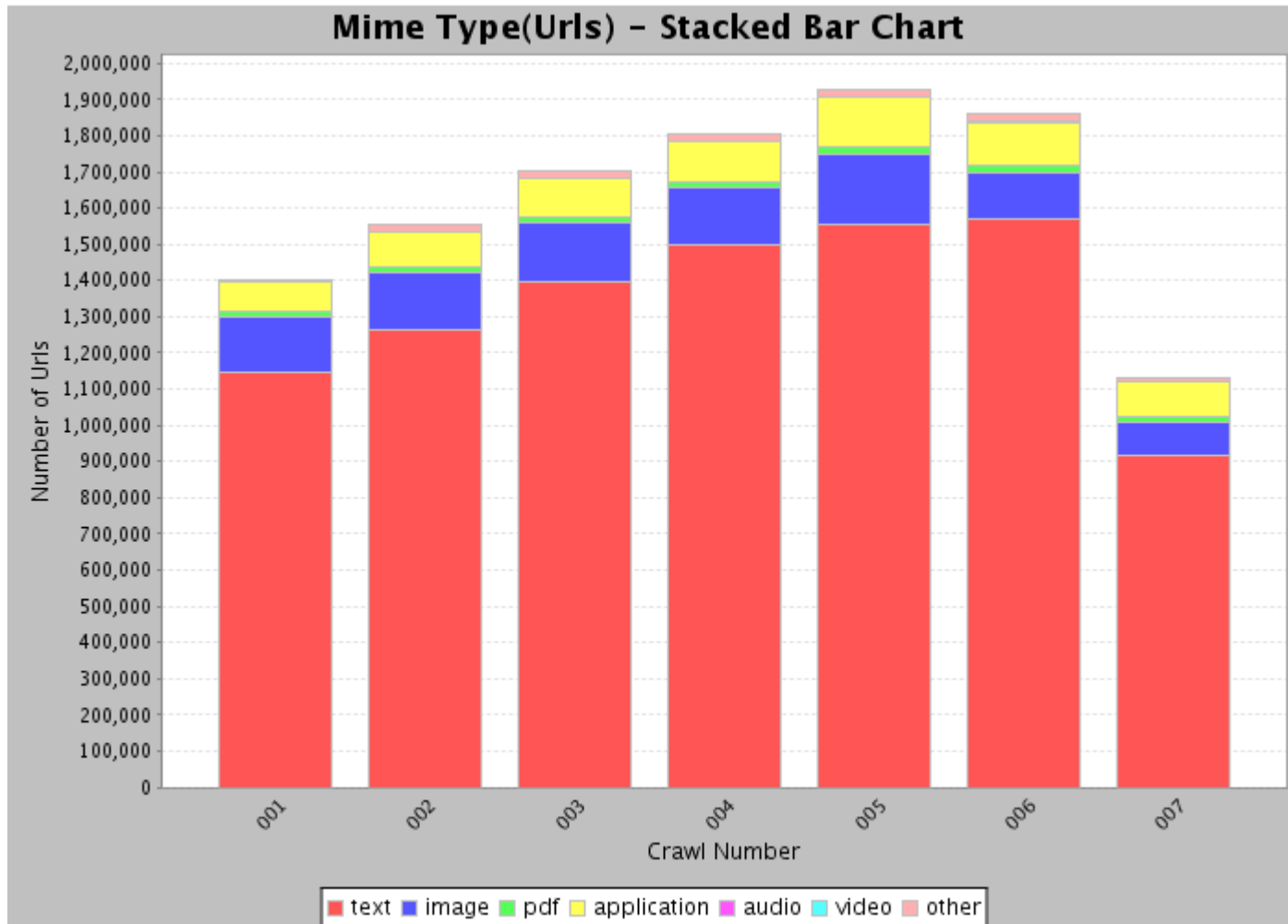
- Sheer mass of data
  - Preservation strategies
    - Migration or Emulation?
    - Does everything in the archive needs preserving?
  - WARC
    - ISO standard-2009
    - Sponsored by the International Internet Preservation Consortium
    - Container for web archive records
-

---

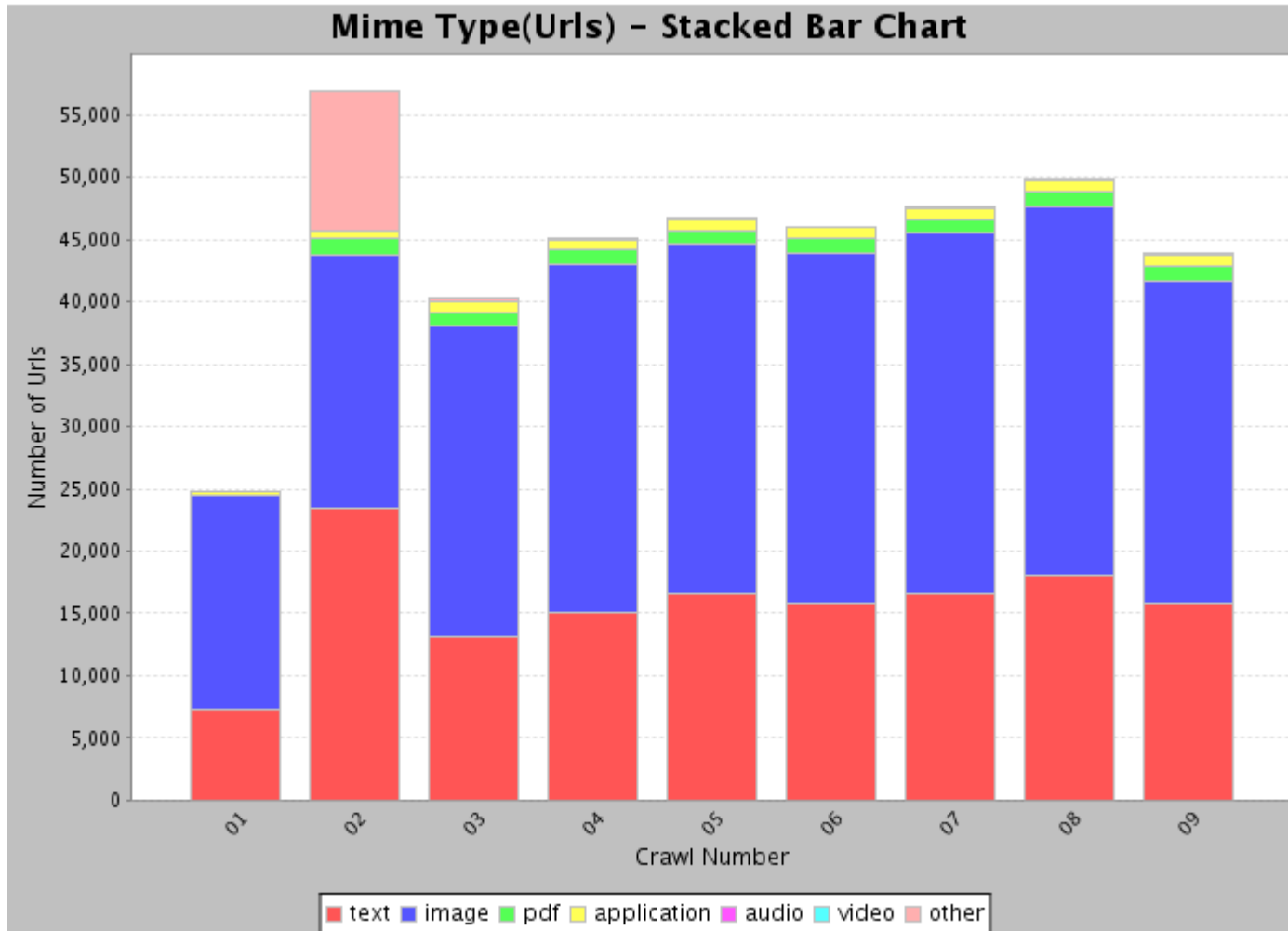
# WARC as a Preservation Tool

- Expected to become the standard way to structure, manage and store large archives.
  - Significant properties of the web archive object.
  - Accommodates secondary content
    - Assigned metadata
    - Duplicate detection events
    - Later day transformations
    - Segmentation of large files
-

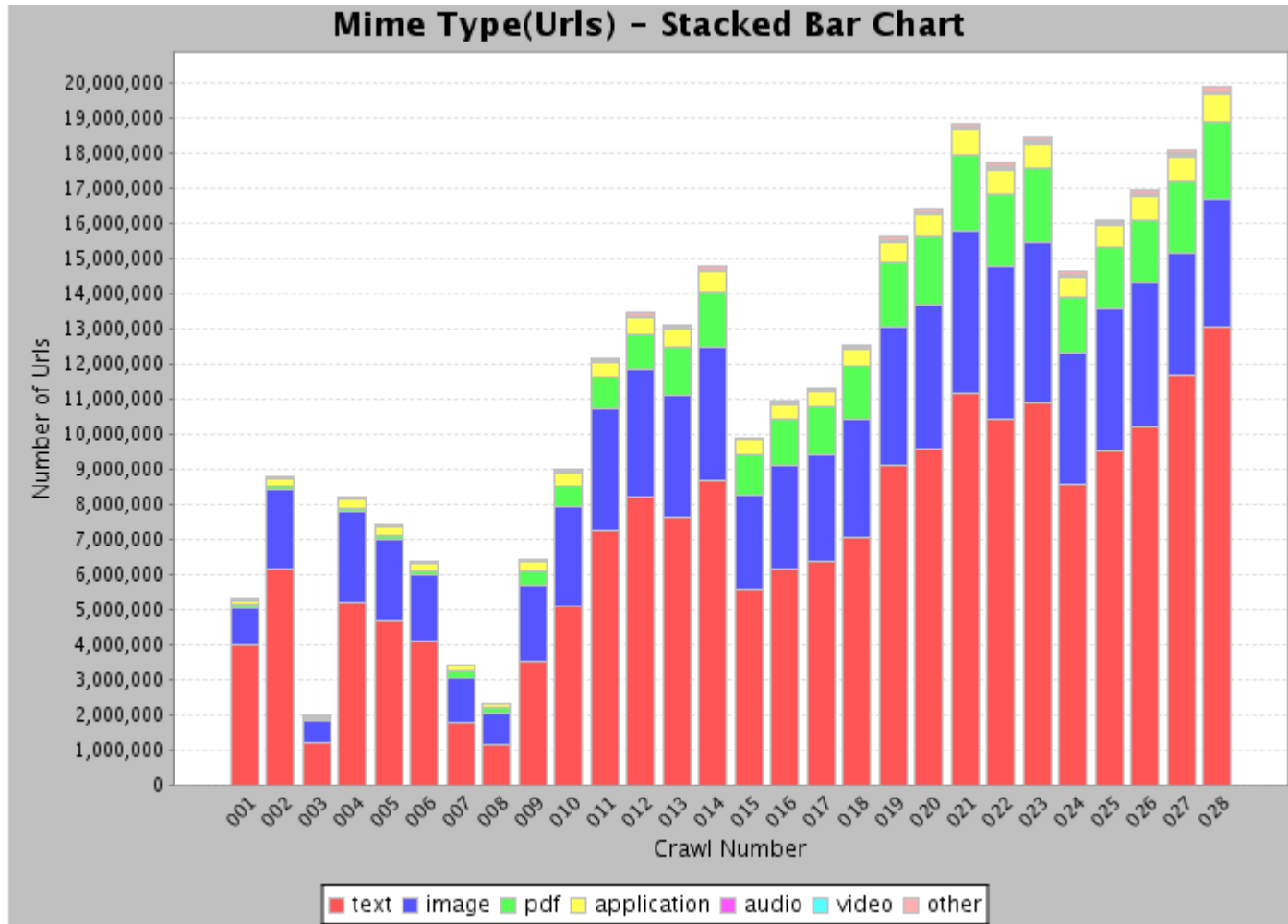
# Legal Blawgs Web Archive



# Prints and Photographs Web Archive



# Primarily Government (State/Local/Federal)



# Deduped ARC Record

Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://webarchives.loc.gov/collections/lcwa0001/20011019014000/http://www.accesshelp.org

DiGiBoard NOMINATE Getting Started Latest Headlines Leaderboard NS2: Niche Social Net... Netpreserve Member... Apache Tomcat/5.5... Home - Internet Arch... caci h

 **LIBRARY OF CONGRESS**

Note: External links, forms and search boxes may not function within this collection

**September 11 Web Archive Collection**  
This is an archived Web site from the Library of Congress  
http://www.accesshelp.org  
Archived: **10/19/2001** at 01:40:00

[◀ Back to previous page](#)      ◀ First (10/19/2001)   ◀ Previous #1 of 7   **Next ▶**   **Last (09/13/2002) ▶**

NEWS\_LOC\_crawl5.20011019013959 102150

---

# Questions

- What kinds of metadata do we need to do to ensure that our web archives are still usable in 10 or more years?
  - What tools are/will be available to help automate the development of metadata?
  - What other kind of metadata should we be preserving to explain our archives?
-