Web Harvesting to Augment Archival Collecting

John P. Rees
Archives and Modern Manuscripts Program
National Library of Medicine
National Institutes of Health
U.S. Department of Health & Human Services
Premise

the web is ephemeral
Archival collections we manage are represented on the web
Archivists should do something to capture and preserve relevant web content as part of collecting mission

Web collecting recently added to NLM Collection Development Policy
M.O.
Use NLM Archive It account to crawl websites related to personal paper and corporate record collections
Use 1st 100 results from a Google search

Or 50

Or whenever good content peters out
Use to augment personal and biographical-like series
Lightly weed out bad links

(don’t have time to deeply investigate every link)
Selection Criteria

– Obituaries, esp. local
– Personal or institutional websites/blogs/tweets
– News sites
– YouTube
– Online publications? Google scholar, paywall content to which you subscribe?
Things to look out for

– Obviously spurious content
– No student projects
– Controversial events -- don’t over-select (be representative but don’t let it cloud larger story)
How John does it
Bernadine Healy
https://en.wikipedia.org/wiki/Bernadine_Healy

Bernadine Patricia Healy (August 4, 1944 – August 6, 2011) was an American physician, cardiologist, academic, and first female National Institutes of Health Director. Wikipedia
You've visited this page 2 times. Last visit: 8/27/17

Dr. Bernadine P. Healy, Influential Health Administrator, Dies at 67...
Aug 8, 2011 - Dr. Bernadine P. Healy, the first woman to lead the National Institutes of Health and the first physician to lead the American Red Cross until she ...

Dec 22, 2001 - Deborah Sontag article on American Red Cross's internal conflict and downfall of its president. Dr Bernadine Healy; notes conflict stems from ...

Bernadine Healy, M.D. | National Institutes of Health (NIH)
Dr. Bernadine Healy became the 13th NIH director in April 1991, appointed by President George H.W. Bush. Shortly after her appointment, she launched the NIH ...

Changing the Face of Medicine | BernadineHealy
https://medicinet.mnh.nig.gov/physicians/biography_145.html
Cardiologist Bernadine Healy was a physician, educator, and health administrator who was the first woman to lead the National Institutes of Health (NIH).
Save the underlying HTML
Short Unix script to grab just the URLs and dedupe

grep "<a href=" healy.htm|sed "s/<a href/
\\n<a href/>g"|sed
's/&quot;/"&gt;&lt;/a\\n/2'|grep href|sort|uniq >
healy-out.txt
Throw the list into Archive It and crawl
Bernadine Healy Papers

Collected by: National Library of Medicine

Archived date: Oct 21, 2017

Description: Archive-It web archive, top results for “Bernadine Healy” Google search conducted by HMD staff as an effort to augment her personal papers collection


Date: Nov 3, 2011

Group: Archives and Manuscript Collections

Collector: National Library of Medicine, NLM

Narrow Your Results

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

Page 1 of 1 (66 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Bernadine Healy, MD (1944–2011) - AHRPAHRP
URL: http://ahrp.org/bernadine-healy-md/
Captured once on Nov 3, 2011

Title: Red Cross pay to ex-president, aide scrutinized - tribunedigital-baltimore
No Captums were found for this URL

Title: For First Time, Red Cross Picks A Doctor As President - tribunedigital-chicagotribune
URL: http://articles.chicagotribune.com/1999-07-08
Add Archive IT Collection URL to finding aid
Red Cross Head Stepping Down

Healy reportedly tabbed for NIH post

WASHINGTON -- Dr. Bernadine Healy of the Cleveland Clinic Foundation apparently has been picked by Health and Human Services Secretary Louis Sullivan to fill the long-vacant job of director of the National Institutes of Health, officials said Monday.

'As best as I can tell, she is Dr. Sullivan's nominee,' said William Raub, acting NIH director, prior to a news conference on women's health issues at the NIH's Bethesda, Md., campus.
Bernadine Healy

From Wikipedia, the free encyclopedia

Bernadine Patricia Healy (August 4, 1936 – August 6, 2011) was an American physician, cardiologist, academic, and first female National Institutes of Health (NIH) Director. She was a professor of medicine at Johns Hopkins University, chair of the Research Division of Cleveland Clinic, professor and dean of the College of Medicine and Public Health at the Ohio State University, and served as president of the American Red Cross and the American Heart Association. She was also a health editor and commentator for U.S. News & World Report. She was a well-known commentator in the media on health issues.71

Contents

• Early years and family
• Affiliations
• Cleveland Clinic
• American Red Cross
• Government service
• 1. President, American Heart Association
• 2. National Institutes of Health
• 2. Women's Health Initiative
• 3. U.S. Senate candidate
• 4. Ohio State University
• 5. Advisory boards
• 6. Press
• 7. Family
• 8. Death
• 9. References
• 10. External links

Dr. Bernadine Healy talks about vaccines and autism...or does she?

Vaccines and autism: publicity of the topic just got a “shot in the arm” this weekend with a story on Dateline. As part of the story, Dr. Bernadine Healy was interviewed. Dr. Healy has called for more research into the proposed vaccine-autism link. She has some good credentials (former head of the National Institutes of Health).

Take a look at what she had to say:

http://www.msnbc.msn.com/id/32413001/tp/3254905492584965

And words are due to Bernadine Healy, unless noted, and taken about the facts.
Thanks!

reesj@nlm.nih.gov