

APACHE
LUCENE
EUROCON



Archive-It: Scaling Beyond a Billion Archival Web-pages

Aaron Binns, Internet Archive
aaron@archive.org, 2011-10-19

Presented by

lucid
IMAGINATION

Lucene

Apache
Solr

My Background

- Aaron Binns (aaron@archive.org)
- Internet Archive
- Senior Software Engineer
- Full-text search & cool stuff
 - Full-text search
 - Hadoop
 - “Big Data”
- <http://github.com/aaronbinns>

Internet Archive

- Universal access to all knowledge
- <http://archive.org>
- Founded 1996
- 501(c)(3) non-profit org.
- Digital Library
- San Francisco, CA, USA
- 7+ PB of publicly accessible digital materials
 - **Web archive**
 - **Books, music, video, etc.**



- <http://web.archive.org>
- 165,000,000,000+ archived web pages
 - **HTML**
 - **Images**
 - **CSS**
 - **JavaScript**
 - **Multimedia**
- 1996-today

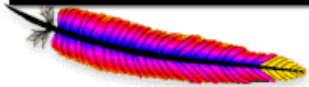
Jakarta Lucene - Overview x +

web.archive.org/web/20011120102851/http://jakarta.apache.org/lucene/docs/index.html

INTERNET ARCHIVE
WayBackMachine BETA

229 captures
20 Nov 01 - 14 May 11

Go OCT NOV FEB Close X
2000 20 2001 2003 Help ?



The Jakarta Project

<http://jakarta.apache.org>



About

- [Overview](#)
- [Powered by Lucene](#)
- [Who We Are](#)
- [Mailing Lists](#)
- [Bugs](#)

Jakarta Lucene

Jakarta Lucene is a high-performance, full-featured text search engine written entirely in java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

Jakarta Lucene is an open source project available for [free download](#) from Apache Jakarta. Please use the links on the left to access Lucene.

Documentation

- [FAQ](#)
- [Articles](#)
- [Javadoc](#)

Lucene News

Lucene v1.02 released - This release repackages Lucene as product of the Apache Software Foundation. Download it [here](#).

Download

- [Binaries](#)
- [Source Code](#)
- [CVS Repositories](#)

Lucene Joins Jakarta - The Lucene Team is happy to announce that Lucene is now part of a member of the Apache Jakarta Project. This move will help Lucene continue to grow, and enhance its position as the leading server-side searching solution for java.

Jakarta

- [Get Involved](#)
- [Acknowledgements](#)

About Apache Jakarta

The goal of the [Apache Jakarta Project](#) is to provide commercial-quality server

<http://archive-it.org>

- Subscription web archiving service
 - Select websites to harvest, frequency, depth
 - Crawling/Harvesting
 - Wayback
 - Full-text search
- Customers
 - Public, State & University Libraries
 - Local governments
 - Museums
 - Non-Governmental Organizations (NGOs)

Collections & Documents

- Collection
 - Web harvest configuration
 - URLs to crawl
 - Frequency & depth
 - Set of documents archived
 - Access via Wayback Machine
 - Full-text search
- Document
 - Unique version of a URL
 - “Text” documents: HTML, PDF, Office, etc.

Archive-It: Collection

COLLECTION: North Carolina State Government Web Site Archive

Partner: [North Carolina State Archives and State Library of North Carolina](#)

Crawling Activity: 2005 - present

Topics: Government - US States

Videos Archived: [list videos](#)



The North Carolina State Government Web site Archive contains content from the Web sites of North Carolina State Government Agencies, Occupational Licensing Boards, and Commissions. This collection includes many different types of records, in a variety of file formats (such as text-based information, images, and video and audio files) including meeting minutes, policies, and agency publications. The State Library of North Carolina and the North Carolina State Archives have collaborated on this initiative to identify, capture, and make accessible state government information.

 [Advanced Search](#)

Prefer to search by URL? Enter URLs Here...

- <http://142.165.148.220/>
- <http://149.168.6.211/sn/spiders/>
- <http://204.211.89.70/> - **Division of Environmental Health - Home**
This page includes information on the Division of Environmental Health, its employees and activities.
- <http://abcs.ncpublicschools.org/> - **ABCs accountability model**
The abcpublicschools.org site is the web interface that contains information regarding North Carolina public schools, the Federal "No Child Left Behind Act" passed in 2002. It also details how often and in what grade children are tested on subject and include a link to the Proctor's guide. This site specifically addresses the accountability in public schools. In addition, there are links to reports that document what schools in each county have been designated a school of excellence and therefore qualifies for more federal funding.
- <http://agronomy.agr.state.nc.us/> - **North Carolina Department of Agriculture & Consumer Services Agronomic Division - Public Access Labo**
North Carolina Department of Agriculture Agronomic Division site providing users with information on soil testing, heavy metals, pH, soluble salts, soil acidity, soil fertility, plant nutrients, water quality, environmental protection, soil sampling, soil test reports, agronomic practices, stewardship, and online reports.
- <http://apnepblog.blogspot.com/>
- <http://apps.dot.state.nc.us/controlpanel/news> - **North Carolina Department of Transportation Public Information Office Releases**
This is the Public Information Office's news release web page. This page has excerpts and links to current news related to the Department of Transportation.
- <http://ayp.ncpublicschools.org/> - **Adequate Yearly Progress (AYP) Reports - North Carolina Schools**
This page contains links to the annual AYP reports, school improvement reports, and performance composite information.

Archive-It: Wayback

wayback.archive-it.org/194/*/http://www.healthycarolinians.org/



Search Results for Jan 1, 1996 - Dec 31, 2011

00	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
es	2 pages	16 pages	9 pages	9 pages	60 pages	27 pages	27 pages	11 pages	6 pages	4 pages	4 pages
	Apr 19, 2001 *	Jul 27, 2002 *	Mar 20, 2003	Feb 7, 2004	Feb 3, 2005 *	Jan 4, 2006	Jan 5, 2007	Jan 18, 2008	Jan 16, 2009 *	Apr 16, 2010 *	Jan 25, 2011 *
	May 17, 2001 *	Sep 14, 2002	Mar 25, 2003	Apr 10, 2004 *	Feb 5, 2005	Jan 19, 2006	Jan 10, 2007	Jan 29, 2008	Mar 21, 2009 *	Jun 18, 2010 *	Apr 6, 2011 *
		Sep 14, 2002	May 25, 2003 *	Apr 20, 2004	Feb 12, 2005	Jan 25, 2006	Jan 14, 2007	Feb 16, 2008	May 15, 2009 *	Aug 20, 2010 *	Jun 4, 2011 *
		Sep 16, 2002	May 26, 2003	May 18, 2004	Mar 8, 2005	Jan 27, 2006 *	Jan 19, 2007	Mar 15, 2008 *	Jul 24, 2009 *	Nov 20, 2010 *	Sep 15, 2011 *
		Sep 18, 2002	Jun 10, 2003	Jun 10, 2004	Apr 8, 2005	Feb 1, 2006	Jan 29, 2007	Apr 14, 2008	Sep 18, 2009 *		
		Sep 27, 2002	Jul 31, 2003	Jun 11, 2004	May 19, 2005	Feb 4, 2006	Feb 2, 2007	Apr 18, 2008	Nov 20, 2009 *		
		Oct 8, 2002	Aug 3, 2003	Sep 24, 2004 *	Jun 18, 2005 *	Feb 18, 2006	Feb 5, 2007	Jun 13, 2008			
		Oct 17, 2002	Oct 29, 2003	Nov 9, 2004 *	Aug 8, 2005 *	Feb 19, 2006	Mar 4, 2007	Jun 18, 2008			
		Oct 28, 2002	Nov 22, 2003	Nov 25, 2004	Aug 18, 2005	Apr 20, 2006 *	Apr 27, 2007 *	Aug 15, 2008 *			
		Nov 4, 2002			Aug 26, 2005	Apr 24, 2006	Jun 8, 2007	Oct 24, 2008 *			
		Nov 14, 2002			Sep 8, 2005	Apr 28, 2006	Jun 15, 2007	Dec 19, 2008 *			
		Nov 19, 2002			Sep 9, 2005	Apr 28, 2006	Jun 15, 2007 *				
		Nov 25, 2002			Sep 21, 2005	May 28, 2006	Jun 17, 2007 *				
		Nov 27, 2002			Sep 21, 2005	Jun 1, 2006	Aug 26, 2007 *				
		Nov 28, 2002			Sep 22, 2005	Jun 15, 2006	Sep 13, 2007				
		Dec 2, 2002			Sep 22, 2005	Jul 5, 2006 *	Sep 17, 2007				
					Sep 23, 2005	Jul 12, 2006 *	Sep 26, 2007				
					Sep 23, 2005	Aug 10, 2006 *	Sep 29, 2007				



Archive-It: Replay



HEALTHY CAROLINIANS

HEALTHY CAROLINIANS

Welcome and Good Evening

WHAT'S NEW?

- Introducing the [Community Micro Grants Project](#): Available to community-based organizations
- The [11th Annual Healthy Carolinians Conference](#): Conference Objectives, Agenda, Keynote Speaker, Pre-Registration and more!
- [May 2002](#) Healthy Carolinians E-News!

508 BOBBY APPROVED

Best Viewed at 800 X 600 Resolution on Netscape

NC 2010 Health Objectives
Certification Process
Governor's Task Force
Office of Healthy Carolinians
County Profiles
Community Assessment
Training & Resources
Conference Information
Web Links

July 27, 2002



contact us | site map | search

HEALTHY CAROLINIANS

HEALTHY CAROLINIANS

❖ HOME

❖ ABOUT US

❖ GOVERNOR'S TASK FORCE

❖ COUNTY PROFILES

❖ HEALTHY CAROLINIANS CONFERENCE

❖ HEALTHY NORTH CAROLINIA 2020 OBJECTIVES

❖ COMMUNITY HEALTH ASSESSMENT

Welcome to Healthy Carolinians!

We help communities improve the health of their website is a guide, tool and resource for communities who want to improve health and safety in their town and county. Join us and begin making a difference in your community.

On this site you can learn how your local community can participate in a community health assessment. Your community group can be certified, or recertified, as a Healthy Carolinians partnership. Find out what's new and improve the health and safety of your county or town.

countries. And be inspired by some of our success stories

Sept 15, 2011

Archive-It: Search

Full Text Results

Results **1 - 20** of about **10,000** for **healthy carolinians** found in 0.3 seconds

[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [next»](#) [last»»](#)

[Healthy Carolinians](#)

Healthy Carolinians contact us | site map | search GO HOME ABOUT US Who we are What we... **CAROLINIANS** CONFERENCE 2011 2010 2009 Past Conferences **HEALTHY** NORTH CAROLINIA 2020 OBJECTIVES COMMUNITY... Collaborative CERTIFICATION/ RECERTIFICATION PROCESS How to become a **Healthy Carolinians** Partnership... NC SOPHE and Governor's Task Force for **Healthy Carolinians** Conference more > 2010 2010 **Healthy Carolinians** Conference Presentations more > 2009 2009 **Healthy Carolinians** Conference Presentations more > 2007 - 2008 2007 - 2008 **Healthy Carolinians** Conference Presentations more >

Welcome to **Healthy Carolinians!** We help communities...

text/html - 21.1 KB - crawled once Sep 15, 2011

<http://www.healthycarolinians.org/> - [more results from healthycarolinians.org](#)

[NC DHHS Release: Report: North Carolinians need to eat more fruits and vegetables](#)

NC DHHS Release: Report: North **Carolinians** need to eat more fruits and vegetables Skip all navigation Skip to page navigation Search DHHS: Home | A-Z Site Map | Divisions & Offices | About Us | Contacts Facts and Figures Department-Level Statistics and Publications News and Notices Press... Report: North **Carolinians** need to eat more fruits and vegetables North Carolina already at work to improve **healthy** eating opportunities Release Date: October 7, 2009 Contact: Carol Schriber, 919-733-9190 RALEIGH – North **Carolinians** are not eating their fruits and vegetables, according to a recent... policies that can be put in place or improved to promote **healthy** eating among our residents," Engel..., to promote increased opportunities for **healthy** eating and physical activity wherever people live...

text/html - 16.4 KB - crawled 5 times between Nov 20, 2010 and Sep 15, 2011 - [list all captures](#)

<http://www.ncdhhs.gov/pressrel/2009/2009-10-7-fruitsveggies.htm> - [more results from ncdhhs.gov](#)

[Gov. Perdue Urges North Carolinians to Honor Workers While Enjoying A Safe and Happy Holiday Weekend](#)

Gov. Perdue Urges North **Carolinians** to Honor Workers While Enjoying A Safe and Happy Holiday Weekend this site nc.gov Home Contact Your State Government Governor's Office Phone Numbers, Staff Listing, Physical Address and Regional Offices Request a Certificate, Proclamation or Award Visiting the State Capitol North Carolina's Governors Visiting the Executive Mansion Cabinet Agencies Council of... Release 9/3/2011 Contact: Chris Mackey RALEIGH 919-733-5612 Gov. Perdue Urges North **Carolinians** to... following statement regarding the holiday weekend: "I urge all North **Carolinians** on this Labor Day weekend... greater aid than the prayers and helping hands of their fellow North **Carolinians**. "I wish everyone a safe, **healthy** and happy Labor Day weekend and encourage the people of North Carolina to enjoy the...

text/html - 23.9 KB - crawled once Sep 15, 2011

<http://www.governor.state.nc.us/NewsItems/PressReleaseDetail.aspx?newsItemID=2028> - [more results from governor.state.nc.us](#)

[North Carolinians Learn about their Risk for Diabetes Alert Day - Blog by April Reese Acting Branch](#)

North **Carolinians** Learn about their Risk for Diabetes Alert Day - Blog by April Reese Acting Branch Head for Diabetes Preve... | Facebook Email Password Keep me logged in Forgot your password? Sign Up Facebook helps you connect and share with the people in your life. Bev Perdue's Notes (0) Notes About Bev Perdue (0) Subscribe Bev Perdue's Notes North **Carolinians** Learn about their Risk for Diabetes Alert Day - Blog by April Reese Acting Branch

Archive-It: Search



Full Text Results

Results 1 - 20 of about 6,000 for **healthy carolinians** from the host **healthycarolinians.org** found in 0.3 seconds [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [next»](#) [last»»](#)

Healthy Carolinians

Healthy Carolinians contact us | site map | search GO HOME ABOUT US Who we are What we... **CAROLINIANS** CONFERENCE 2011 2010 2009 Past Conferences **HEALTHY** NORTH CAROLINIA 2020 OBJECTIVES COMMUNITY... Collaborative CERTIFICATION/ RECERTIFICATION PROCESS How to become a **Healthy Carolinians** Partnership... NC SOPHE and Governor's Task Force for **Healthy Carolinians** Conference more > 2010 2010 **Healthy Carolinians** Conference Presentations more > 2009 2009 **Healthy Carolinians** Conference Presentations more > 2007 - 2008 2007 - 2008 **Healthy Carolinians** Conference Presentations more > Welcome to **Healthy Carolinians!** We help communities...
text/html - 21.1 KB - crawled once Sep 15, 2011
<http://www.healthycarolinians.org/>

Healthy Carolinians

Healthy Carolinians contact us | site map | search GO HOME ABOUT US Who we are What we... Health Disparities **Healthy** People 2010 **Healthy Carolinians** 2010 Strengths of a Diverse Population.../ RECERTIFICATION PROCESS How to become a **Healthy Carolinians** Partnership Certification Standards Benefits... SUCCESS STORIES LINKS AND RESOURCES 9/30/2010 - 10/1/2010 18th Annual **Healthy Carolinians** Conference, Greensboro, NC more > 2009 2009 **Healthy Carolinians...** **Healthy Carolinians** Conference Presentations more > Calendar... certified, or recertified, as a local **Healthy Carolinians** partnership. Find out what's happening to improve... Assessment **Healthy Carolinians** Conference Assessment is a vital part of improving the health of...
text/html - 30.8 KB - crawled once Jan 25, 2011
<http://www.healthycarolinians.org/>

Healthy Carolinians

Healthy Carolinians contact us | site map | search GO HOME ABOUT US Who we are What we... Health Disparities **Healthy** People 2010 **Healthy Carolinians** 2010 Strengths of a Diverse Population.../ RECERTIFICATION PROCESS How to become a **Healthy Carolinians** Partnership Certification Standards Benefits Timeline... Resources SUCCESS STORIES LINKS AND RESOURCES 9/30/2010 - 10/1/2010 18th Annual **Healthy Carolinians...** - 2008 **Healthy Carolinians** Conference Presentations more... **Healthy Carolinians!** We help communities improve the health of their citizens. This website is a guide... certified, or recertified, as a local **Healthy Carolinians** partnership. Find out what's happening to... Health Assessment **Healthy Carolinians** Conference Assessment is a vital part of improving the health of...
text/html - 30.8 KB - crawled once Nov 20, 2010
<http://www.healthycarolinians.org/>

Healthy Carolinians

Healthy Carolinians contact us | site map | search GO HOME ABOUT US Who we are What we... **CAROLINIANS** CONFERENCE 2010 2009 Past Conferences **HEALTHY** NORTH CAROLINIA 2020 OBJECTIVES COMMUNITY HEALTH... assessment action plans. 9/30/2010 - 10/1/2010 18th Annual **Healthy Carolinians** Conference, Greensboro, NC more > 2009 2009 **Healthy Carolinians** Conference Presentations

Challenges and....Solutions?

- Scale
- *Archival* web search != web search
- Document formats
 - HTML (1996....2011)
 - PDF, Office, text, etc.
- English, Français, Español, 漢字, ...
- Diversity
- Time

Scale

- 200+ customers
- 2,272 collections
 - Largest: 33,470,659 documents
 - 24 collections, 10,000,000+ docs
 - 250 collections, 1,000,000+ docs
- Total:
 - 1,375,473,187 *unique* documents

Scale...each day

- 30-40 simultaneous crawls/harvests
- ~150GB of data: HTML, images, media
- ~1.3 million new *unique* documents
 - New URLs never seen before
 - New versions of URLs
- ~1.3 million *updates*
 - Documents unchanged
 - New crawl dates

Architecture

- Offline indexing
 - 10 dedicated indexing machines
 - ~10% of collections per machine
 - Add new documents
 - Update existing documents with new dates
 - 1CPU x 2core, 4GB RAM, 3x2TB disk
- Search service
 - 11 machines: 1 master, 10 slaves
 - ~10% of collections per slave
 - 1 collection → 1 Lucene index
 - 1CPU x 2core, 8GB RAM, 3x2TB disk

Diversity

[Apache Lucene - Overview](#)

[lucene.apache.org](#) > Lucene - Traducir esta página

Apache **Lucene**(TM) is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application ...

[Lucene Java Documentation - Apache Lucene - Downloads and ... - Features](#)

[Welcome to Apache Lucene!](#)

[lucene.apache.org/](#) - Traducir esta página

Search engine library with many features including fast indexing, ranked ...

[Apache Lucene.Net](#)

[incubator.apache.org/lucene.net/](#) - Traducir esta página

23 Jul 2011 – Port of the Java **Lucene** search engine to the C# and .NET ...

[+](#) [Mostrar más resultados de apache.org](#)

[Lucene - Wikipedia, la enciclopedia libre](#)

[es.wikipedia.org/wiki/Lucene](#)

Saltar a: navegación, búsqueda. **Lucene** es un API de código abierto para recuperación de información, originalmente implementada en Java por Doug Cutting. ...

[Lucene - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Lucene](#) - Traducir esta página

Apache **Lucene** is a free/open source information retrieval software library ...

[+](#) [Mostrar más resultados de wikipedia.org](#)

[Apache Lucene - Dos Ideas](#)

[www.dosideas.com/wiki/Apache_Lucene](#)

Apache **Lucene** es un motor de búsqueda de texto, escrito en Java. **Lucene** es una excelente opción para cualquier aplicación que requiera búsqueda de datos ...

[Introducción a Lucene](#)

[www.adictosaltrabajo.com/tutoriales/tutoriales.php?pagina=lucene...](#)

17 Ago 2008 – **Lucene** es una librería (no un producto) open source (con licencia de ... **Lucene** va mucho más allá que las búsquedas en BD a través de ...

Diversity

Full Text Results

Results **1 - 20** of about **10,000** for **healthy carolinians** found in 0.3 seconds

[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [next»](#) [last»»](#)

[Healthy Carolinians](#)

Healthy Carolinians contact us | site map | search GO HOME ABOUT US Who we are What we... **CAROLINIANS** CONFERENCE 2011 2010 2009 Past Conferences **HEALTHY** NORTH CAROLINIA 2020 OBJECTIVES COMMUNITY... Collaborative CERTIFICATION/ RECERTIFICATION PROCESS How to become a **Healthy Carolinians** Partnership... NC SOPHE and Governor's Task Force for **Healthy Carolinians** Conference more > 2010 2010 **Healthy Carolinians** Conference Presentations more > 2009 2009 **Healthy Carolinians** Conference Presentations more > 2007 - 2008 2007 - 2008 **Healthy Carolinians** Conference Presentations more >

Welcome to **Healthy Carolinians!** We help communities...

text/html - 21.1 KB - crawled once Sep 15, 2011

<http://www.healthycarolinians.org/> - [more results from healthycarolinians.org](#)

[NC DHHS Release: Report: North Carolinians need to eat more fruits and vegetables](#)

NC DHHS Release: Report: North **Carolinians** need to eat more fruits and vegetables Skip all navigation Skip to page navigation Search DHHS: Home | A-Z Site Map | Divisions & Offices | About Us | Contacts Facts and Figures Department-Level Statistics and Publications News and Notices Press... Report: North **Carolinians** need to eat more fruits and vegetables North Carolina already at work to improve **healthy** eating opportunities Release Date: October 7, 2009 Contact: Carol Schriber, 919-733-9190 RALEIGH – North **Carolinians** are not eating their fruits and vegetables, according to a recent... policies that can be put in place or improved to promote **healthy** eating among our residents," Engel..., to promote increased opportunities for **healthy** eating and physical activity wherever people live...

text/html - 16.4 KB - crawled 5 times between Nov 20, 2010 and Sep 15, 2011 - [list all captures](#)

<http://www.ncdhhs.gov/pressrel/2009/2009-10-7-fruitsveggies.htm> - [more results from ncdhhs.gov](#)

[Gov. Perdue Urges North Carolinians to Honor Workers While Enjoying A Safe and Happy Holiday Weekend](#)

Gov. Perdue Urges North **Carolinians** to Honor Workers While Enjoying A Safe and Happy Holiday Weekend this site nc.gov Home Contact Your State Government Governor's Office Phone Numbers, Staff Listing, Physical Address and Regional Offices Request a Certificate, Proclamation or Award Visiting the State Capitol North Carolina's Governors Visiting the Executive Mansion Cabinet Agencies Council of... Release 9/3/2011 Contact: Chris Mackey RALEIGH 919-733-5612 Gov. Perdue Urges North **Carolinians** to... following statement regarding the holiday weekend: "I urge all North **Carolinians** on this Labor Day weekend... greater aid than the prayers and helping hands of their fellow North **Carolinians**. "I wish everyone a safe, **healthy** and happy Labor Day weekend and encourage the people of North Carolina to enjoy the...

text/html - 23.9 KB - crawled once Sep 15, 2011

<http://www.governor.state.nc.us/NewsItems/PressReleaseDetail.aspx?newsitemID=2028> - [more results from governor.state.nc.us](#)

[North Carolinians Learn about their Risk for Diabetes Alert Day - Blog by April Reese Acting Branch](#)

North **Carolinians** Learn about their Risk for Diabetes Alert Day - Blog by April Reese Acting Branch Head for Diabetes Preve... | Facebook Email Password Keep me logged in Forgot your password? Sign Up Facebook helps you connect and share with the people in your life. Bev Perdue's Notes (0) Notes About Bev Perdue (0) Subscribe Bev Perdue's Notes North **Carolinians** Learn about their Risk for Diabetes Alert Day - Blog by April Reese Acting Branch

Diversity

Full Text Results

Results 1 - 20 of about 6,000 for **healthy carolinians** from the host **healthycarolinians.org** found in 0.3 seconds [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [next»](#) [last»»»](#)

Healthy Carolinians

Healthy Carolinians contact us | site map | search GO HOME ABOUT US Who we are What we... **CAROLINIANS** CONFERENCE 2011 2010 2009 Past Conferences **HEALTHY** NORTH CAROLINIA 2020 OBJECTIVES COMMUNITY... Collaborative CERTIFICATION/ RECERTIFICATION PROCESS How to become a **Healthy Carolinians** Partnership... NC SOPHE and Governor's Task Force for **Healthy Carolinians** Conference more > 2010 2010 **Healthy Carolinians** Conference Presentations more > 2009 2009 **Healthy Carolinians** Conference Presentations more > 2007 - 2008 2007 - 2008 **Healthy Carolinians** Conference Presentations more > Welcome to **Healthy Carolinians!** We help communities...
text/html - 21.1 KB - crawled once Sep 15, 2011
<http://www.healthycarolinians.org/>

Healthy Carolinians

Healthy Carolinians contact us | site map | search GO HOME ABOUT US Who we are What we... Health Disparities **Healthy** People 2010 **Healthy Carolinians** 2010 Strengths of a Diverse Population.../ RECERTIFICATION PROCESS How to become a **Healthy Carolinians** Partnership Certification Standards Benefits... SUCCESS STORIES LINKS AND RESOURCES 9/30/2010 - 10/1/2010 18th Annual **Healthy Carolinians** Conference, Greensboro, NC more > 2009 2009 **Healthy Carolinians**... **Healthy Carolinians** Conference Presentations more > Calendar... certified, or recertified, as a local **Healthy Carolinians** partnership. Find out what's happening to improve... Assessment **Healthy Carolinians** Conference Assessment is a vital part of improving the health of...
text/html - 30.8 KB - crawled once Jan 25, 2011
<http://www.healthycarolinians.org/>

Healthy Carolinians

Healthy Carolinians contact us | site map | search GO HOME ABOUT US Who we are What we... Health Disparities **Healthy** People 2010 **Healthy Carolinians** 2010 Strengths of a Diverse Population.../ RECERTIFICATION PROCESS How to become a **Healthy Carolinians** Partnership Certification Standards Benefits Timeline... Resources SUCCESS STORIES LINKS AND RESOURCES 9/30/2010 - 10/1/2010 18th Annual **Healthy Carolinians**... - 2008 **Healthy Carolinians** Conference Presentations more... **Healthy Carolinians!** We help communities improve the health of their citizens. This website is a guide... certified, or recertified, as a local **Healthy Carolinians** partnership. Find out what's happening to... Health Assessment **Healthy Carolinians** Conference Assessment is a vital part of improving the health of...
text/html - 30.8 KB - crawled once Nov 20, 2010
<http://www.healthycarolinians.org/>

Healthy Carolinians

Healthy Carolinians contact us | site map | search GO HOME ABOUT US Who we are What we... **CAROLINIANS** CONFERENCE 2010 2009 Past Conferences **HEALTHY** NORTH CAROLINIA 2020 OBJECTIVES COMMUNITY HEALTH... assessment action plans. 9/30/2010 - 10/1/2010 18th Annual **Healthy Carolinians** Conference, Greensboro, NC more > 2009 2009 **Healthy Carolinians** Conference Presentations

Field Collapsing / Grouping

- Applied to web documents
 - “Give me the best 1-2 hits from a site”
- Lucene
 - Grouping contrib package
- Solr
 - Field Collapsing
- What is the performance cost?
- Custom solution

Time

- User experience & understanding
 - *Archival* web search != web search
- Information Architecture
 - Publication date for web pages – difficult
- Temporal diversity
 - Multiple hits per site
 - Multiple *versions* per URL

Time

wayback.archive-it.org/194*/http://www.healthycarolinians.org/



Search Results for Jan 1, 1996 - Dec 31, 2011

00	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
es	2 pages	16 pages	9 pages	9 pages	60 pages	27 pages	27 pages	11 pages	6 pages	4 pages	4 pages
	Apr 19, 2001 *	Jul 27, 2002 *	Mar 20, 2003	Feb 7, 2004	Feb 3, 2005 *	Jan 4, 2006	Jan 5, 2007	Jan 18, 2008	Jan 16, 2009 *	Apr 16, 2010 *	Jan 25, 2011 *
	May 17, 2001 *	Sep 14, 2002	Mar 25, 2003	Apr 10, 2004 *	Feb 5, 2005	Jan 19, 2006	Jan 10, 2007	Jan 29, 2008	Mar 21, 2009 *	Jun 18, 2010 *	Apr 6, 2011 *
		Sep 14, 2002	May 25, 2003 *	Apr 20, 2004	Feb 12, 2005	Jan 25, 2006	Jan 14, 2007	Feb 16, 2008	May 15, 2009 *	Aug 20, 2010 *	Jun 4, 2011 *
		Sep 16, 2002	May 26, 2003	May 18, 2004	Mar 8, 2005	Jan 27, 2006 *	Jan 19, 2007	Mar 15, 2008 *	Jul 24, 2009 *	Nov 20, 2010 *	Sep 15, 2011 *
		Sep 18, 2002	Jun 10, 2003	Jun 10, 2004	Apr 8, 2005	Feb 1, 2006	Jan 29, 2007	Apr 14, 2008	Sep 18, 2009 *		
		Sep 27, 2002	Jul 31, 2003	Jun 11, 2004	May 19, 2005	Feb 4, 2006	Feb 2, 2007	Apr 18, 2008	Nov 20, 2009 *		
		Oct 8, 2002	Aug 3, 2003	Sep 24, 2004 *	Jun 18, 2005 *	Feb 18, 2006	Feb 5, 2007	Jun 13, 2008			
		Oct 17, 2002	Oct 29, 2003	Nov 9, 2004 *	Aug 8, 2005 *	Feb 19, 2006	Mar 4, 2007	Jun 18, 2008			
		Oct 28, 2002	Nov 22, 2003	Nov 25, 2004	Aug 18, 2005	Apr 20, 2006 *	Apr 27, 2007 *	Aug 15, 2008 *			
		Nov 4, 2002			Aug 26, 2005	Apr 24, 2006	Jun 8, 2007	Oct 24, 2008 *			
		Nov 14, 2002			Sep 8, 2005	Apr 28, 2006	Jun 15, 2007	Dec 19, 2008 *			
		Nov 19, 2002			Sep 9, 2005	Apr 28, 2006	Jun 15, 2007 *				
		Nov 25, 2002			Sep 21, 2005	May 28, 2006	Jun 17, 2007 *				
		Nov 27, 2002			Sep 21, 2005	Jun 1, 2006	Aug 26, 2007 *				
		Nov 28, 2002			Sep 22, 2005	Jun 15, 2006	Sep 13, 2007				
		Dec 2, 2002			Sep 22, 2005	Jul 5, 2006 *	Sep 17, 2007				
					Sep 23, 2005	Jul 12, 2006 *	Sep 26, 2007				
					Sep 23, 2005	Aug 10, 2006 *	Sep 29, 2007				



Searching across collections

- Search all collections of a user
- Search arbitrary group of collections
- 1 collection → 1 Lucene index
 - Search 100 collections....
 - Search 100 indexes
- Collections distributed over 10 searchers

Custom Solutions

- Java
- Built on Lucene
- Investigating Solr
 - Capabilities
 - Cost
- Internet Archive
 - Open Source
 - Apache License
 - <http://github.com/aaronbinns>

Custom Solutions: Indexing

- <http://github.com/aaronbinns/jbs>
- Archive-It & other archival web collections
- Hadoop-based, or stand-alone
- Java code with Lucene
 - Hard-coded “schema” for web documents
 - Title, body, keywords, date, mime-type, etc.
 - Link analysis & curation to augment scoring

Custom Solutions: Searching

- <http://github.com/aaronbinns/tnh>
- Custom Java web application with Lucene
- Federated search
 - 1 master, 10 slaves
 - OpenSearch
- Multiple collections & arbitrary grouping
- CollapsingCollector

CollapsingCollector

- <http://github.com/aaronbinns/tnh>
- Extends `Lucene Collector`
- Field cache: “site”
- Retains top N hits per “site”
 - Control N via URL parameter

Web Archives!

- Archive-It
 - <http://archive-it.org/>
- US National Archives
 - <http://webharvest.gov/>
- UK Web Archive
 - <http://www.webarchive.org.uk/>
 - Solr-based
- Web Archive of Catalonia / PADICAT
 - Biblioteca de Catalunya
 - <http://www.padicat.cat/>