Selecting and Managing Content Captured from the Web: Expanding Curatorial Expertise and Skills in Building Library of Congress Web Archives

Abbie Grotke and Janice E. Ruth

DigCCurr2007: An International Symposium in Digital Curation
University of North Carolina, Chapel Hill, NC
April 20, 2007
Web Capture at LC

- **2000 to 2003**
  - MINERVA: event-based, thematic collecting (Elections, September 11, Iraq War, etc.)
  - Collaborations with Internet Archive, Alexa, WebArchivist.org

- **2004 to today**
  - Identify policy issues, establish best practices, build tools (internally and with other international institutions)
  - Broaden expertise and understanding of Web capture (expand from one curator to many)
  - Collect, manage, and sustain at-risk digital content
Selecting and Managing Content Captured from the Web (SMCCW): Major Project Goals

- Define roles and responsibilities for life cycle management of archived Web content (focusing on Plan, Select, and Get life cycle processes)
- Collect at-risk born-digital content
- Document resource requirements and workflow
- Expand staff participation
  - Train larger group of LS staff
  - Learn by doing
Darfur: Bystanders to Genocide
http://www.bystanderstogenocide.com/

This website catalogues the public statements of U.S. politicians on the genocide in Darfur, and compares what they are saying, to what they are actually doing. It is our hope that you will use this information to urge those in public office to take meaningful action to stop the genocide, and to hold your leaders accountable should they choose to do nothing.

Those who do not react to gross human rights violations elsewhere and do not act on behalf of the victims become complicitous because they enable the perpetrator to carry out his violation of human rights. Indifferent and ignorant bystanders, who continue their usual work in a reign of terror, can be seen as perpetrators not only because of their activities but because they enable the perpetrators to carry out their violations. The bystanders’ silence confirms the perpetrators’ faith in what they are doing.

Learn more about the genocide in Darfur here.
LATEST: Humanitarian crisis worsens

In a crisis, humanitarian aid and political settlement are necessary but not sufficient to prevent genocide. People also need protection.

Demonstration on Whitehall, London
12-3pm, 16 December 2005

Darfur deal by end-year possible but difficult - UN GENEVA, Dec 2 (Reuters) - An agreement by the end of the year on a sustained ceasefire in Darfur is "possible but difficult", the United Nations' top envoy in Sudan, Jan Pronk, said on Friday.

BBC Radio 4's Today Programme reports on the latest situation in Darfur and interviews the UK Secretary of State for International Development, Hilary.
Darfur: A Genocide We Can Stop
http://www.darfurgenocide.org/

The Crime of Our New Century...

The Sudanese Government, using Arab "Janjaweed" militias, its air force, and organized starvation, is systematically killing the black Sudanese of Darfur.

The picture of the boy killed (left) shows what is happening in Darfur. Over two and a half million people, driven from their homes, now face death from starvation and disease as the Government and militias attempt to prevent humanitarian aid from reaching them. The same forces have destroyed the people of Darfur's villages and crops, and poisoned their water supplies, and they continue to murder, rape and terrorize.

This site's sole purpose is to try to save lives by helping stop the genocide in Darfur.

It empowers you to take smart, strategic actions to compel those in

Estimated Death Toll in Darfur:

April 20th:

0400000

Email a friend

Click here to tell someone about this site

Message from Bette Midler

Like many of you, I have recently been deeply affected by the terrible suffering of the people of Darfur, Sudan.
Prokudin-Gorskii Images

Color photographs from the Russian Empire taken a century ago (1909-1915).

Sergei Mikhailovich Prokudin-Gorskii was a color photographer before his time, who undertook a photographic survey of the Russian Empire for Tsar Nicholas II. He was able to capture color by taking three pictures of each scene, each with a different red, green or blue color filter. Walter Frankhauser, a photographer contracted by the Library of Congress, manually registered and cleaned up some 120 of the original high-resolution scans, with breathtakingly beautiful results. The results of his effort can be seen at the online-exhibit The Empire That Was Russia.

Using computer-vision technology to automate the registration process, one can now for the first time view almost the entire collection of the Prokudin-Gorskii photographs in color. The color images on these pages were obtained by automatically registering these three pictures to obtain a color image of each scene. The computer program to do this was written in MATLAB by Frank Dellaert using computer-vision technology commonly used in 'mosaicking'. By clicking the links on this page you can view thumbnails of the almost 2000 images purchased by the Library of Congress, and each thumbnail is linked to a larger version of the
Visual Image: Ann Telnaes
http://www.anntelnaes.com/
Visual Image: Rafael Viñoly Architects
http://www.rvapc.com/
# Single Site Selection Guidelines

http://www.loc.gov/staff/webcapture/ls/smccw/single_site/single20siteREVISED0206.wpd

## LS Project: SMCCW

### Single Site

**Collection Title:**
Using sites submitted via TrackER, this group will develop and coordinate a vetting process to determine what types of single-site resources (those not relating to thematic collection activities) to archive.

### Collected Websites

- [Curator Homepage](#)
- [Single Site Content Group Web Archive Data](#)

### Documentation:

- [Collection Specification](#)
- [Permission Plan](#)
- [June 1, 2006 Meeting](#)
- [July 31, 2006 Meeting](#)

---

*Abbie Grothe*
Project Manager

**Voice:** 202.707.2833
**Email:** abgr@loc.gov
Latest News

6/29/06 - Taking Nanotechnology to Market: One Company’s Strategy
Category: Events > Perspectives

What are the challenges of moving nanotechnology from laboratory to store shelves? Featured speaker: Barry Park, Ph.D., Chief Operating Officer, Oxnica.

6/20/2006 - Nanotechnology and Patent Pools: One Stop License Shopping
Category: News

An examination of the viability of using patent pools in nanotechnology as a means to avoid the high cost associated with acquiring numerous licensing agreements needed to create a product.

6/12/2006 - Nanotechnology Risk and Oversight: It’s Your Business!
Category: Events

With Lux Research reporting more than $32 billion in products incorporating emerging nanotechnology sold last year and global nanotechnology R&D spending reaching $9.6 billion annually, governments around the world are looking at ways to ensure workplace, consumer and environmental health and safety.
The primary objective of the MNB shall be to achieve and maintain price stability. Without prejudice to its primary objective, the MNB shall support the economic policy of the Government using the monetary policy instruments at its disposal.

### Current data and statistics

<table>
<thead>
<tr>
<th>Inflation</th>
<th>Central Bank base rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium term target</td>
<td>October 25, 2006</td>
</tr>
<tr>
<td>Actual (November 2006, KSH)</td>
<td>3,0%</td>
</tr>
<tr>
<td>6,4%</td>
<td><strong>8,00%</strong></td>
</tr>
</tbody>
</table>

**Base rate history**

<table>
<thead>
<tr>
<th>Date</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 9, 2007</td>
<td><strong>8,00%</strong></td>
</tr>
<tr>
<td>December 29, 2006</td>
<td>Consolidated balance sheet of MFI's: November 2006</td>
</tr>
</tbody>
</table>

**Currency exchange rates**

<table>
<thead>
<tr>
<th>EUR</th>
<th>USD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EUR 253,33</strong></td>
<td></td>
</tr>
</tbody>
</table>
SMCCW Collections: Number of URLs Selected and Crawled Per Project

- **Manuscript**: 35 selected, 30 crawled
- **Darfur**: 264 selected, 224 crawled
- **Visual Image**: 23 selected, 17 crawled
- **Single Site**: 55 selected, 23 crawled
Workshops and Training

• Workshops
  - Selection
  - Technology (Crawling, scoping, and QR)
  - Copyright and permissions
  - Access tools

• Training
  - Nomination/permissions tool ("Leaderboard")
  - Quality Review - how to evaluate crawled content
Selection Workshop

• How do traditional selection policies and best practices fit Web content?

• What do we collect? What is the scope?
  - Size of collection, start and end dates, frequency, depth, breadth
  - Non-thematic/“Single site” archiving

• How do the technical aspects of the Web affect selection criteria?

• What skills do the curators need to develop, what training is required?
Technology Workshop: Crawling, Scoping, Quality Review, and Other Issues

- Introduction to basic concepts, terms, and definitions
- Overview of existing tools for capturing and viewing Web archives. What are their limitations?
- What does it mean to do quality review on a Web archive? Who reviews what?
- How much information is too much information?
Copyright and Permissions Workshop

• LC does not currently have the legal authority to capture

Permissions planning
  - Roles and responsibilities
    • LS/Recommenders: contact identification and mailing
    • OSI/Web capture team: permissions support and tools
    • Questions from site owners: LS and OSI staff
  - What permission request is sent to what type of site (crawl vs. display)

• Research and Mailing of Permissions
  - Finding reliable contact information harder than expected
  - Significant time commitment for a good response rate
  - Typically 2 to 10 min. per URL, sometimes 30 min. to an hour. Extreme cases sometimes took more research time.
  - Second and third attempts sometimes tried when no response received
Other Training

- Workshop on access tools
- “Lessons Learned” Workshop
- Instruction in using the Leaderboard tool
- Quality review training
Manuscript Division Archive of Organizational Web Sites

- Collection period: April - November 2006
- Frequency: Monthly
- URLs recommended: 35
- URLs collected: 30 (much higher if LWV sites are counted)

Manuscript Stats

- 86% Crawl Granted & Can Show Onsite/Offsite
- 14% Crawl Denied
- 0% No response (not crawled)
MSS Project Timeline

- July 2005 - Team assembled to reflect different sections
- August 2005 - Kick-off meeting of joint OSI-LS project
- Early Fall 2005 - Training, test sites, permissions plan, and selection criteria
- November 2005 - Meeting with chief and specialists
- Early Winter - Permissions plan and selections finalized
- March 2006 - Permission letters mailed
- April 2006 - Crawls begin
- June and August 2006 - Detailed quality review
- November 2006 - Crawls concluded
- December 2006 - Analysis and feedback
Key Tasks

- Appraising and selecting Web sites
- Securing permissions
- Conducting quality review
- These project tasks reflect the Manuscript Division’s existence in a larger institutional setting where it benefits from strong technical support provided by other offices
Knowledge and Skills for Appraising and Selecting Web Sites

- Subject expertise in American history
- Skill at appraising and organizing paper records
- Knowledge of organizational missions and histories
- Experience navigating the Web and evaluating information found there
- Understanding how Web sites are constructed
- Training in selecting seed URLs and recognizing the crawler’s limitations
Appraisal Considerations

• Sparked useful discussions about electronic media among staff
• Decisions tied to existing policies and holdings
• Evaluated potential research value of each site
• Took into account our relationship to the donor
• Analyzed how a Web site complemented, duplicated, or expanded information in the organization's paper-based records at LC
• Considered significance of an organization's activities in relation to the division's overall documentation strategies
New Census Poverty Figures Don't Tell Whole Story for Minorities and Women
The U.S. poverty rate held steady in 2005 - at 12.6 percent - according to new Census Bureau figures released August 29. However, for minorities no change means that millions are still worse off than whites. 2005 was the first year since 1999 that the real median household income increased, but Black and Hispanic households still have the lowest median incomes. And while the poverty rates for... full story | more stories

Federal Government Losing Ground in Hiring Minorities to Senior Positions
More minorities are working in the federal workforce, but they are not among the... full story | more stories

Katrina's One-Year Anniversary Yields Harsh Retrospective
Dawn Peterson, Miss Wheelchair Louisiana 2005, was paralyzed from the waist down... full story | more stories

More on Civil Rights!
LCCR Voting Record 109th Congress

Check it out!
› Take the Pledge to End Poverty in America
› Report: The Faces of Hurricane Katrina (PDF)

The Buzz
The latest in civil rights news!
› NCES Calls for Sticking to the Stats
› With an Unusually Hands-On Role, State Feels Its Way in New Orleans
› Downward Mobility
Welcome to the National Consumers League

October is National Cyber Security Month

Learn to Keep Safe from ID Thieves and Other Ghouls on social networking sites

Let the National Consumers League be your guide to these issues:
- Health
- Fraud
- Money & Credit
- Finances
- Workers' Rights
- Safety & Nutrition
- Food
- Technology
- Utilities
- Privacy
- Other Interests

National Cyber Security Alliance

StaySafeOnline.org

Fraud.org • LifeSmarts • StopChildLabor.org • SOSRx • PhishingInfo.org

NCL 1701 K Street, NW, Suite 1200, Washington DC 20006. 202-835-3323 fax 202-835-0747 info@nclnet.org
MSS: League of Women Voters (National)
http://www.lwv.org/

League Strongly Urges Representatives to Support a Ban on Leadership PACs
The LWVUS joined with other organizations in a letter to U.S. Representatives urging them to publicly support legislative efforts to ban leadership PACs.

League Applauds Senate Vote to Reauthorize the Voting Rights Act
Calls on President Bush to Sign Bill

League Opposes Barriers to Citizen Participation in Voting
The LWVUS sent a letter to the U.S. House of Representatives opposing HR 4844, a bill that would create new barriers to voting by eligible citizens by requiring a photo ID at the polling place and proof of citizenship in the voter registration process.
Welcome to the
League Of Women Voters
Of Philadelphia

Philadelphia's Source for Non-Partisan Election Information

LWV Philadelphia presents the

Philadelphia 2006 Voters Guide

Important information about the 2006 races for US Senate, US House of Representatives, Pennsylvania Governor, Lt. Governor, Pennsylvania Senate (even number districts only) and Pennsylvania House of Representatives.
The American Historical Association (AHA) was founded in 1884 and incorporated by Congress in 1889 to serve the broad field of history. It encompasses every historical period and geographical area and serves professional historians in all areas of employment. Read more...
Welcome to the American Studies Association Homepage

Jump Menu (Select a Section and Press "GO")
A-Z Index
Search the ASA site on Crossroads:

- What is the American Studies Association?
- Vote in the 2006 ASA Election (Click Here).
- Awards and Prizes 2006
  Contribute to the Anzaldúa Award for Independent Scholars
- ASA Minority Scholars' Committee Homepage
MSS: Vietnam Veterans Memorial Fund
http://www.vvmf.org/

May 22 — Annual Memorial Day Observance at The Wall
Thousands of veterans, families, friends and visitors will come to the Vietnam Veterans Memorial on May 29 for the annual Memorial Day Observance at The Wall at 1:00 p.m. During the ceremony, the four names recently added to The Wall will be "touched." The featured speaker will be Willard T. Frank and singer/conductor Jimmy Fortune will perform "More Than a Name on The Wall." Read the Press Release, View Webcast

May 15 — Names to be Added to The Wall
The names of four American servicemen will be inscribed on the black granite walls of the Vietnam Veterans Memorial next week. Work will begin on Tuesday, May 15, and a ceremony will be held for the press to witness the addition of one of the names on Thursday, May 18. These additions will bring the number of names on The Wall to 56,223 men and women who were killed in Vietnam or remain missing in action. Read the Press Release

May 10 — Mother's Day at The Wall
Mothers whose children served and sacrificed in the Vietnam War and other military conflicts will be honored at the Vietnam Veterans Memorial this Sunday. The American Gold Star Mothers, American War Mothers and the Girl Scouts of the Nation's Capital will be participating in the ceremony scheduled to begin at 9 a.m. Read the Media Advisory

April 6 — American Heroes to be Honored Posthumously
Eighty-six American heroes will be honored posthumously during the 8th annual "Memory Day" ceremony to be held at the Vietnam Veterans Memorial April 15. The Memorial Fund has partnered with TVWorldwide.com to webcast many of the major events, including ceremonies at The Wall and teachers' conferences.

Offerings at The Wall
Receive this beautifully photographed book of remembrances left at The Wall with your generous donation of $75 or more. (This offer only applies to online donations and while supplies last.)

Make a donation
More about the book

The Wall That Heals
The Wall That Heals will stop in Marion, IL (1100 Block of S. Ariens Street) June 1-4. View the 2006 Schedule to see if the exhibit will be stopping near you this year.
Historical Society of the District of Columbia Circuit

Available Histories

- Hon. Patrick J. Attidge
- Hon. Robert H. Bork
- Hon. James L. Buckley
- Edmund D. Campbell, Esq.
- Lloyd N. Cutler
- Hon. Thomas A. Flannery
- Milton V. Freeman, Esq.
- Warner W. Gardner, Esq.
- Hon. Oliver Gasch
- Hon. Gerhard A. Gesell
- Hon. Harold H. Greene
- Daniel M. Gribbon, Esq.
- Erwin N. Criswell, Esq.
- Charles A. Horsky, Esq.
- Hon. George E. MacKinnon
- Hon. Ahmen I. Mikva
- Herbert J. Miller, Jr., Esq.
- Bernard I. Nordlinger, Esq.

Oral History Program

The Society's Oral History Program preserves the recollections of judges, lawyers, and others who have helped to shape the history of the Circuit. Recorded interviews, later transcribed, capture the voices, inflections, insights, and stories of the very people who know most intimately how the courts have functioned and developed over the years. Photographs of many of the subjects are also available.

Conducted by volunteer attorneys trained by the Program, oral history interviews provide anecdotal information about family relationships and backgrounds, personal observations about career histories and interactions with colleagues, reflections on the times, comments on law practice and significant court decisions, the administration of justice, life in the courts of the D.C. Circuit, and lessons learned from lifetimes in the law.

An invaluable resource for historians and students of legal process and the administration of justice, the completed oral histories are maintained in the Judges' Library at the E. Barrett Prettyman U.S. Courthouse, the Manuscript Reading Room of the Library of Congress, and the Historical Society of Washington, D.C. (See list of completed histories.) Some of the histories are available on this site. Additional histories will be added to the site as they are completed.
The Association for Diplomatic Studies and Training (ADST) is a private nonprofit organization founded in 1986. Located in an historic Sears-style cottage at the State Department's George P. Shultz National Foreign Affairs Training Center in Arlington, Virginia, ADST advances understanding of American diplomacy and supports training of foreign affairs personnel at the NFATC's Foreign Service Institute (FSI) through a variety of programs and activities. This
The Frederick A. Cook Society

"Frederick Cook remains as one of the most perceptive as well as the most vilified of polar explorers...Cook had made a discovery that no amount of humiliation could take from him and one that only the most fortunate and observant of arctic explorers ever learned."

Robert McGhee,  
"The Last Imaginary Place: A Human History of the Arctic World"  
Oxford University Press, 2005

Dr. McGhee is curator of Arctic Archeology at the Canadian Museum of Civilization and Massey Medalist of the Royal Canadian Geographical Society

"If some insist that history must be served, let them remember that the more we come to understand the physical geography of the Arctic Ocean, the more we recognize the phenomena described by Cook but unknown at the time of his journey."

Dr. Walter A. Wood,  
President of the American Geographical Society (1963)

"There is no ground
Sigmund Freud's (1856-1939) discoveries concerning the unconscious mind have had a major influence on Western thought and have permeated contemporary culture. His development of psychoanalysis contributed an essential method for the understanding, treatment and research of psychological disturbance.

The Sigmund Freud Archives, Inc. is an entirely independent organization, founded in 1951. It is dedicated to collecting, conserving, collating and making available for scholarly use all of Sigmund Freud’s psychoanalytic and personal papers, his correspondence, photos, records, memorabilia, etc. These documents are protected and preserved at the United States Library of Congress, in the Freud Collection, established with the collaboration and donation of Sigmund Freud Archives. The Sigmund Freud Archives has a policy of de-restriction, except for ethical and legal constraints and most of the collection is now accessible. All documents are released unaltered save for the deletion of patients' names to preserve anonymity and confidentiality. All interested persons may apply to the Library of Congress for permission to see non-restricted documents on the basis of equal access.

Papers of Anna Freud and other eminent psychoanalysts are available in
FEATURED ARTICLE

NSA Secret Wiretap Program Held Unconstitutional

Aug. 17, 2006
NACDL Press Release

Washington, D.C. — As plaintiffs in the lawsuit seeking to enjoin the government’s warrantless wiretapping program, the National Association of Criminal Defense Lawyers (NACDL) and its members are grateful for today’s courageous decision by U.S. Judge Anna Biggs Taylor. Judge Taylor held the secret “terrorist surveillance program” (TSP) unconstitutional under the First and Fourth Amendments and separation of powers doctrine, and unlawful under the federal Wiretap Act (Title III) and the Foreign Intelligence Surveillance Act (FISA).

→ Read More

NEWS OF INTEREST

The National Endowment for Democracy

Supporting freedom around the world

The Endowment is guided by the belief that freedom is a universal human aspiration that can be realized through the development of democratic institutions, procedures, and values. Governed by an independent, nonpartisan board of directors, the NED makes hundreds of grants each year to support pro-democracy groups in Africa, Asia, Central and Eastern Europe, Eurasia, Latin America, and the Middle East.

Spotlight:

Esther Dyson on "A Virtualist Tours the World: The Internet and the Emergence of Democracy"

Esther Dyson, longtime analyst of the IT and Internet worlds and author of the blog "Release 0.9," addressed the topic, "A Virtualist Tours the World: The Internet and the Emergence of Democracy," at the New York Democracy Forum on Nov. 27.

Saad Eddin Ibrahim on Toward Islamic Democracies

Toward Islamic Democracies was the topic for the Third Annual Seymour Martin Lipset Lecture on Democracy in the World, delivered by Dr. Saad Eddin Ibrahim on Nov 1. [more...]
  ○ View this event

Dr. Jorge Castañeda on Democratization in the Western Hemisphere: Prospects and Challenges

Jorge Castañeda, former Foreign Minister of Mexico and renowned public
Permissions Process

- Mailed personalized form letters to 33 of the 35 organizations; 20 responded positively after receiving the first letter
- Eventually received permission from 28 of the 33 we contacted by letter and 2 whom we contacted by form email
- Total of 30 of 35 sites or 86.6% responded; compares favorably to typical 30% rate
- Sending individualized permission requests was time-consuming and likely not scalable, but important for us as a donor relations tool
Conducting Quality Review

- Core project team received training but not the subject specialists
- Crawls captured both more and less than expected
- Crawling instructions cannot easily exclude parts of a Web site
- Common problems were missing text, broken links, links that go outside the crawled version to the live Web, and content captured but not accessible via the Wayback Machine
Improving Staff Involvement in QR

- Confirmed the need for quality review to be undertaken within the division as well as by the technical support team
- Questions remain about how best to implement and routinize that review
- Possible reasons for uneven results include unclear expectations; lack of time; lack of proofreading or editorial skills; limited technical understanding; and greater acceptance of errors
- Checklists, more training, and single point person needed
Conclusions and Next Steps

- Despite lingering questions about costs, workflow, and technology, interest remains high and benefits are notable.
- SMCCW project successfully increased expertise and understanding of Web capture throughout the Library.
- Significant progress made in resolving curation issues in three stages of the digital life cycle—PLAN, GET, and SELECT.
- More work needed on the other stages, particularly DESCRIBE, MAKE AVAILABLE, and SUSTAIN.
- Likely expansion of Single Site project will provide opportunities for more staff training, skill development, and study of additional issues.
For More Information

• Public Web Capture site: http://www.loc.gov/webcapture/

Abbie Grotke
Digital Media Project Coordinator
Office of Strategic Initiatives
abgr@loc.gov

Janice E. Ruth
Manuscript Specialist in American Women’s History
Manuscript Division
jrut@loc.gov