The NDSA Content Working Group's National Agenda Digital Content Area Discussions: Web and Social Media

Abbie Grotke, Library of Congress
NDSA Content Working Group Co-Chair
@agrotke
abgr@loc.gov

About the National Agenda for Digital Stewardship

- The *National Agenda for Digital Stewardship* annually integrates the perspective of dozens of experts and hundreds of institutions to provide funders and executive decision – makers insight into emerging technological trends, gaps in digital stewardship capacity, and key areas for funding, research and development to ensure that today's valuable digital content remains accessible and comprehensible in the future, supporting a thriving economy, a robust democracy, and a rich cultural heritage.

- http://www.digitalpreservation.gov/ndsa/nationalagenda/index.html

Content Areas and Challenges

- Areas of content featured in 2014 report:
  - **Web and Social Media** (today!)
  - Electronic Records (January 8, 2014)
  - Moving Image and Recorded Sound (March 5, 2014)
  - Research Data (April 2, 2014)

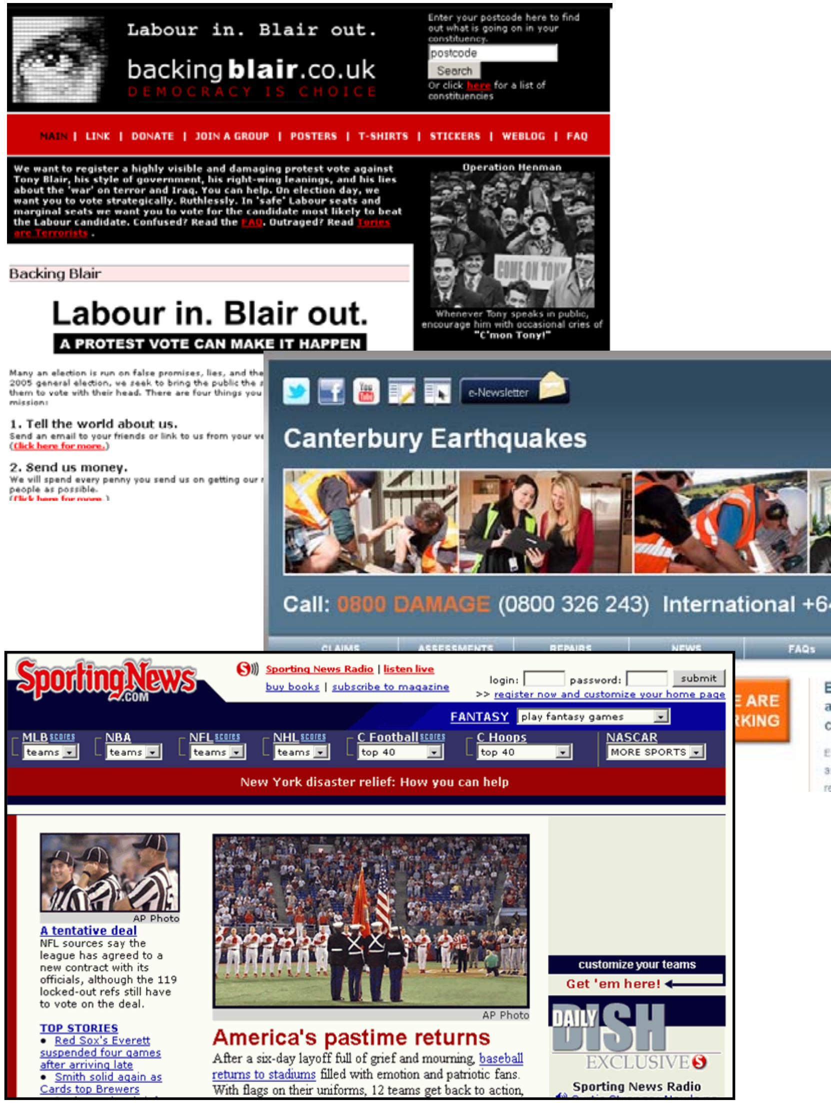
- Major challenges throughout:
  - Size of data requiring preservation
  - Selection of content when the totality cannot be preserved
  - Determining how to store and migrate content to ensure long-term preservation.

Web and Social Media

Why Preserve the Web and Social Media?

Information published on the Web today will be the primary resources for future researchers. Web archiving documents:

- > events unfolding on the web
- > changes in web technologies, design, functionality
- > changes and record of governments
- > the creative output of citizens
- > entire domains of some countries to meet legal deposit mandates
- > websites of institutions and companies for compliance and records management



Labour in. Blair out.

backing blair.co.uk

DEMOCRACY IS CHOICE

Enter your postcode here to find out what is going on in your constituency.

Or click here for a list of constituencies

We want to register a highly visible and damaging protest vote against Tony Blair, his style of government, his right-wing leanings, and his lies about the 'war' on terror and Iraq. You can help. On election day, we want you to vote strategically. Ruthlessly. In 'safe' Labour seats and marginal seats we want you to vote for the candidate most likely to beat the Labour candidate. Confused? Read the FAQ. Outraged? Read Tories are Terrorists .

Operation Henman

Whenever Tony speaks in public, encourage him with occasional cries of "C'mon Tony!"

Backing Blair

Labour in. Blair out.

A PROTEST VOTE CAN MAKE IT HAPPEN

Many an election is run on false promises, lies, and the
2005 general election, we seek to bring the public the
them to vote with their head. There are four things you
mission:

1. Tell the world about us.
Send an email to your friends or link to us from your we
(Click here for more.)

2. Send us money.
We will spend every penny you send us on getting our
people as possible.
(Click here for more.)

Canterbury Earthquakes

Call: 0800 DAMAGE (0800 326 243) International +6

New York disaster relief: How you can help

A tentative deal
NFL sources say the league has agreed to a new contract with its officials, although the 119 locked-out refs still have to vote on the deal.

America's pastime returns
After a six-day layoff full of grief and mourning, baseball returns to stadiums filled with emotion and patriotic fans. With flags on their uniforms, 12 teams get back to action,

AP Photo

AP Photo

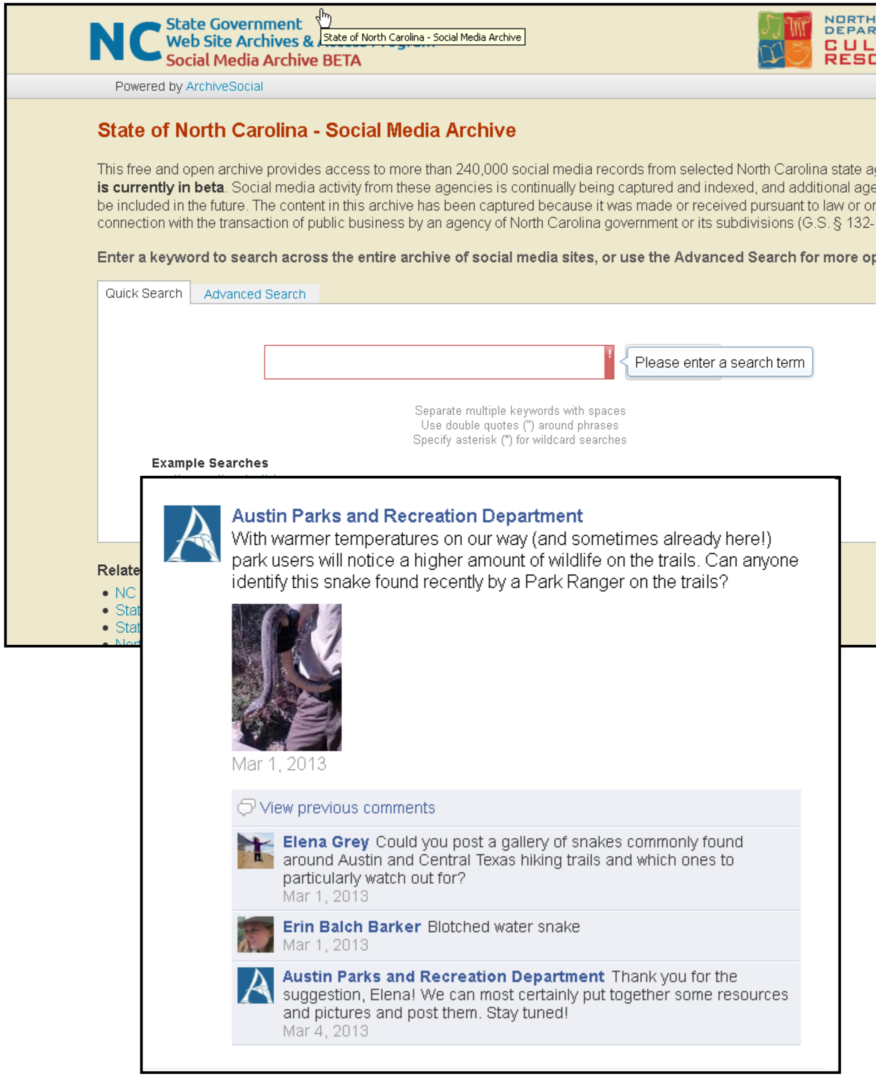
TOP STORIES
- Red Sox's Everett suspended four games after arriving late
- Smith solid again as Cards top Brewers

Social Media Preservation

- society is turning to social media as a primary method of communication and creative expression
- social media is supplementing letters, journals, serial publications and other sources routinely collected by research libraries
- services hosting this content don't have preservation on the mind - changes they make in how they present content can upset the preservation process
- technology to allow for scholarship access to large data sets is lagging behind technology for creating and distributing such data
- this represents a new kind of collecting for many institutions:
  - are they federal or otherwise official records?
  - what is the value of this content?
  - how much can we preserve?
  - are there privacy concerns?

Who is preserving the web and social media?

- National Libraries and Archives
- Colleges and Universities
- Research Institutions
- Museums and Art Libraries
- Government Agencies
- Public Libraries
- Corporations
- Site owners (Personal Digital Archiving)



NC State Government Web Site Archives &
Social Media Archive BETA

State of North Carolina - Social Media Archive

Powered by ArchiveSocial

State of North Carolina - Social Media Archive

This free and open archive provides access to more than 240,000 social media records from selected North Carolina state a
is currently in beta. Social media activity from these agencies is continually being captured and indexed, and additional age
be included in the future. The content in this archive has been captured because it was made or received pursuant to law or or
connection with the transaction of public business by an agency of North Carolina government or its subdivisions (G.S. § 132-

Enter a keyword to search across the entire archive of social media sites, or use the Advanced Search for more op

Separate multiple keywords with spaces
Use double quotes (") around phrases
Specify asterisk (*) for wildcard searches

Example Searches

Relate
- NC
- Stat
- Stat

Austin Parks and Recreation Department With warmer temperatures on our way (and sometimes already here!) park users will notice a higher amount of wildlife on the trails. Can anyone identify this snake found recently by a Park Ranger on the trails?

Mar 1, 2013

Elena Grey Could you post a gallery of snakes commonly found around Austin and Central Texas hiking trails and which ones to particularly watch out for?
Mar 1, 2013

Erin Balch Barker Blotched water snake
Mar 1, 2013

Austin Parks and Recreation Department Thank you for the suggestion, Elena! We can most certainly put together some resources and pictures and post them. Stay tuned!
Mar 4, 2013

47 member organizations:

national libraries & archives,
universities, service providers

harvesting, preservation, &
access working groups

IIPC
netpreserve.org

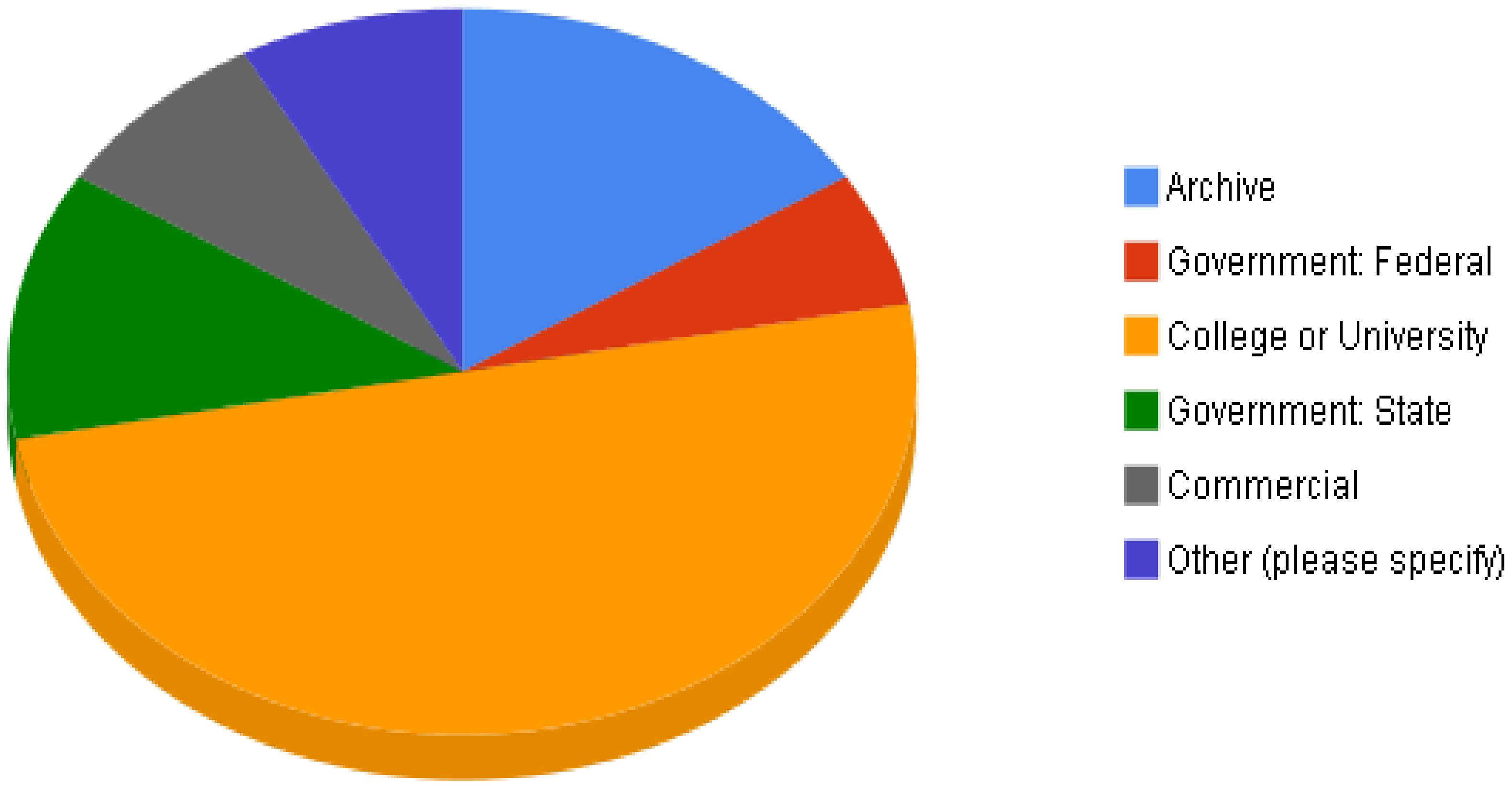
INTERNATIONAL
INTERNET
PRESERVATION
CONSORTIUM

training & educational programs

collaborative collections & access

http://netpreserve.org/

NDSA Web Archiving Survey (as of Nov 5)

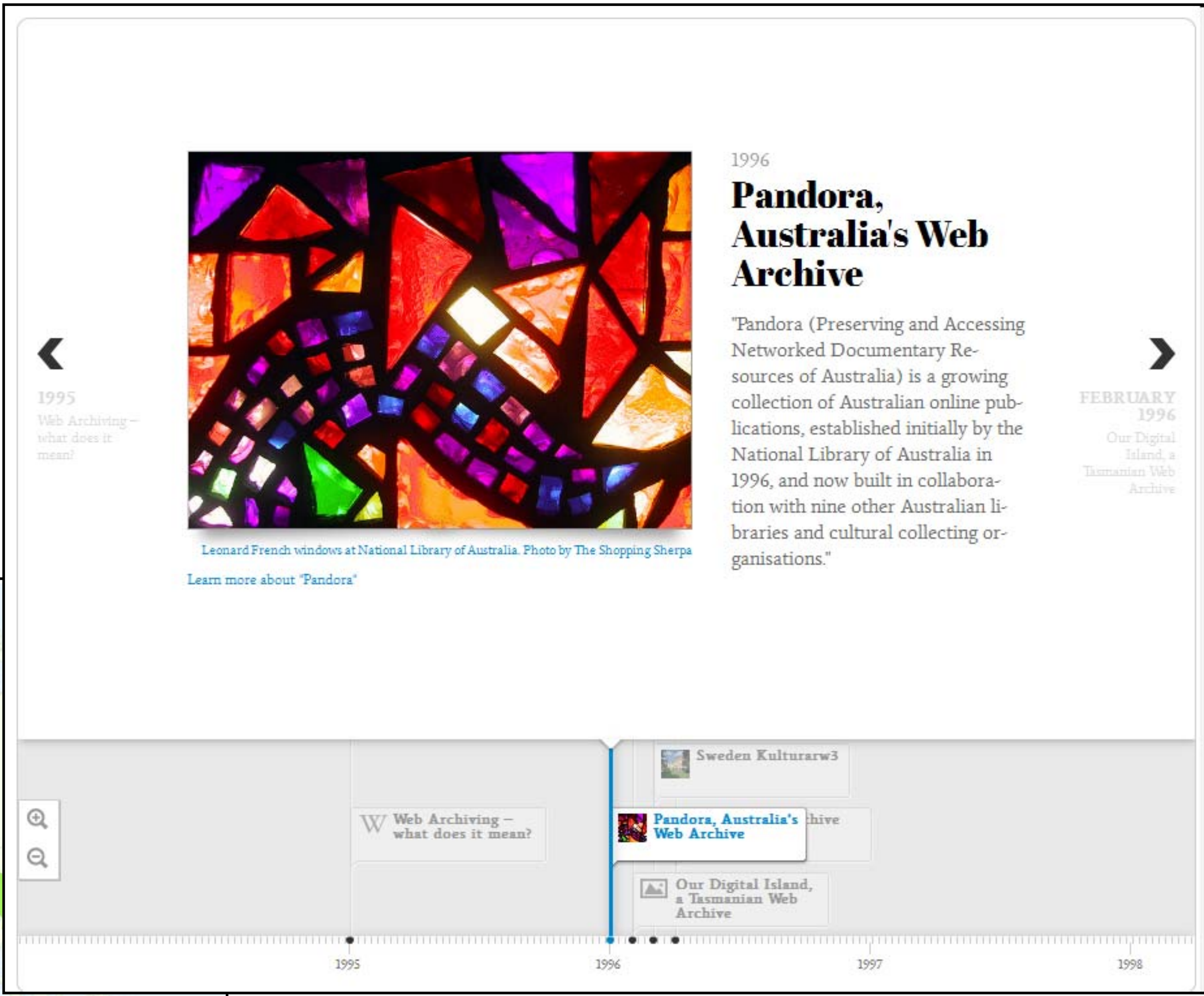


webarchivists

The Web-Archives Timeline

how people around the world are archiving the Internet since 1996

timeline.webarchivists.org


1996

Pandora, Australia's Web Archive

"Pandora (Preserving and Accessing Networked Documentary Resources of Australia) is a growing collection of Australian online publications, established initially by the National Library of Australia in 1996, and now built in collaboration with nine other Australian libraries and cultural collecting organisations."

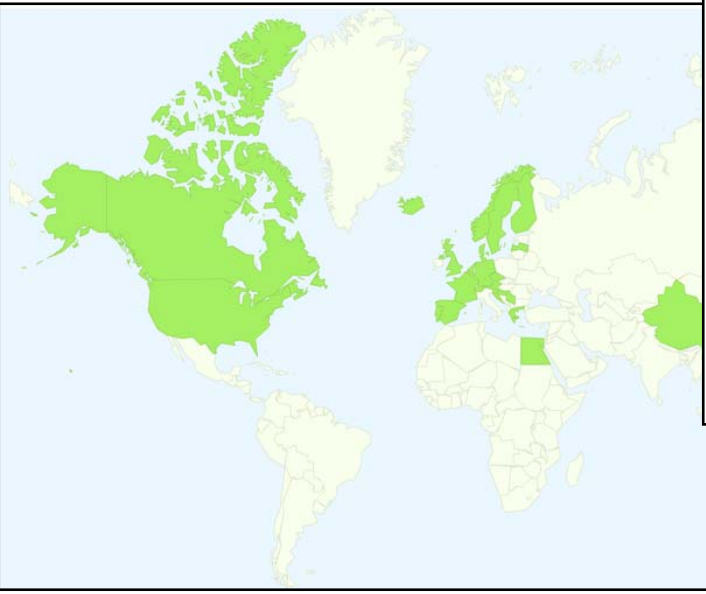
Leonard French windows at National Library of Australia. Photo by The Shopping Sherpa

Learn more about "Pandora"

1995 Web Archiving – what does it mean?

FEBRUARY 1996 Our Digital Island, a Tasmanian Web Archive

Sweden Kulturarw3
Web Archiving – what does it mean?
Pandora, Australia's Web Archive
Our Digital Island, a Tasmanian Web Archive
1995
1996
1997
1998



http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives

Objectives and Approaches

- “Broad” or “global” web (Internet Archive)
- Domain Crawling
  - National Libraries tasked with preserving entire domains
  - Unites States efforts to archive federal government domain (.gov)
- Event or Thematic
  - National, Regional, Local Elections
  - Global events
  - Tragedies and Disasters
- Selective/Representative
  - “top 100” sites in a particular topic
  - News sites

What are some of the challenges?

Ethical and Social Challenges

- With so much content being produced in every part of the globe, who is responsible for preserving it?
- How do we select what gets preserved from the mass?
  - Collection/Selection policies help institutions articulate goals, set boundaries
  - Comprehensive collecting on any given topic nearly impossible – is a "representative" collection good enough?
  - Duplication in collecting – not necessarily bad
    - Though how do we know what others are preserving?

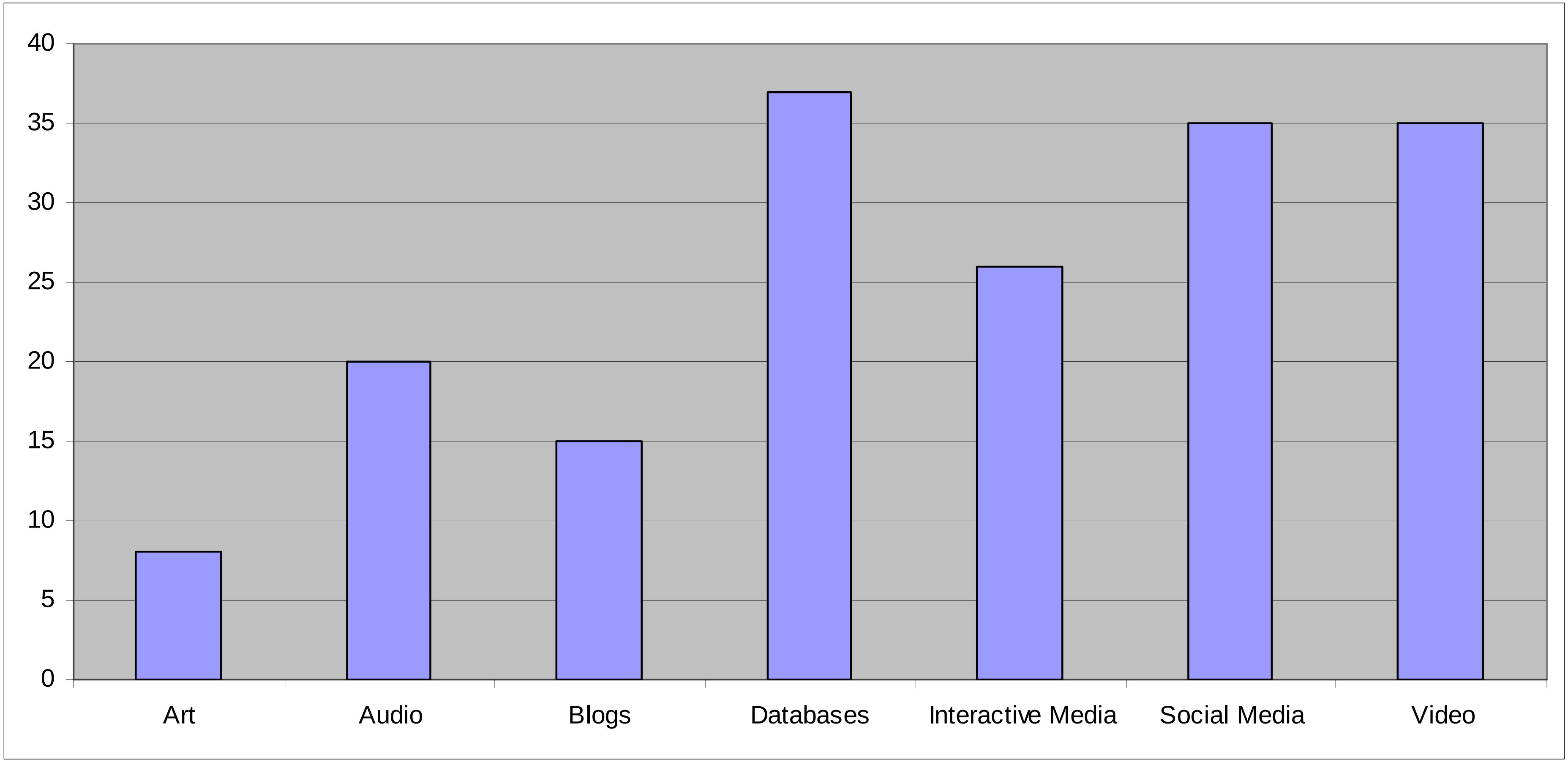
Ethical and Social Challenges: Privacy Concerns

- How do we balance our archiving and research interests in capturing and preserving the content creation and connections with the creator's right to privacy?
- Does it matter if the original content was published on the open web?
  - Challenges for an individual desiring to define and sustain a particular level and type of privacy on the "live web" have been well-documented.
  - "Right to forget" laws in some countries
- Possible research/analysis area: Can we/should we apply existing privacy and personal security-related practices from banks, government agencies to digital preservation?

Technical Challenges

- A constant challenge for the crawling and access tool technologies to keep up with what’s happening on the web
- It is tricky and sometimes impossible to archive certain types of web content
- Requires constant monitoring: A site that archived perfectly yesterday, might not after today’s redesign
- And… sometimes content can be archived, but the access tools cannot display the archived content.

Technical Challenges (2013 NDSA survey)



Legal Challenges

- Web archivists across the globe face different legal frameworks:
  - some are waiting for legislation
  - others have legislation that covers web archiving
  - others have legal doctrines such as fair use or legal deposit that permit or mandate web archiving
  - many follow a permissions-based or notification approach, in absence of legislation, or if the legal frameworks are unclear.

http://www.netpreserve.org/web-archiving/legal-issues

Legal challenges for those asking permissions

- Lack of response from site owners.
- Patchy, unbalanced collections as a result of permissions not granted.
- Determining whether 3rd party rights need to be secured.
- The tremendous effort required to contact site owners and notify or obtain permission can sometimes overwhelm staff resources.
- Risk assessments and fair use analysis may allow some organizations to do more, however some are hesitant to go down this path and instead take a more cautious approach.

Legal issues: Miscellaneous Approaches

- Access Embargos
- «Respecting» (or not) robots.txt
  - Robots.txt is a file that websites use to provide instructions to crawlers.
  - Respecting robots.txt can interfere with archiving in a number of ways:
    - Entire sites can be blocked with robots.txt, or specific parts of sites.
    - Sometimes style sheets and images will be blocked, elements that are important when you are trying to document the look and feel of a website.
  - Some organizations obey robots.txt except when it comes to inline images and stylesheets, so the website is better represented. Others who are seeking permission bypass the robots.txt so that the sites archived are as complete as possible.

More information

- National Agenda http://www.digitalpreservation.gov/ndsa/nationalagenda/index.html
- 2011 NDSA Web Archiving survey results http://digitalpreservation.gov/ndsa/working_groups/documents/ndsa_web_archiving_survey_report_2012.pdf (2013 to come!)
- IIPC website: http://www.netpreserve.org
- UNT/IIPC 2013 Bibliography http://digital.library.unt.edu/ark:/67531/metadc172362/
- Archive-IT Life Cycle Model

http://archive-it.org/static/files/archiveit_life_cycle_model.pdf

- SAA Web Archiving Roundtable

http://webarchivingrt.wordpress.com/