STANFORD UNIVERSITY LIBRARIES

Boiling the Ocean, Together: Web Archive Collection Development in a Global Context

Nicholas Taylor
Web Archiving Service Manager
Digital Library Systems and Services

Chalk Talk
May 12, 2014

Distributed

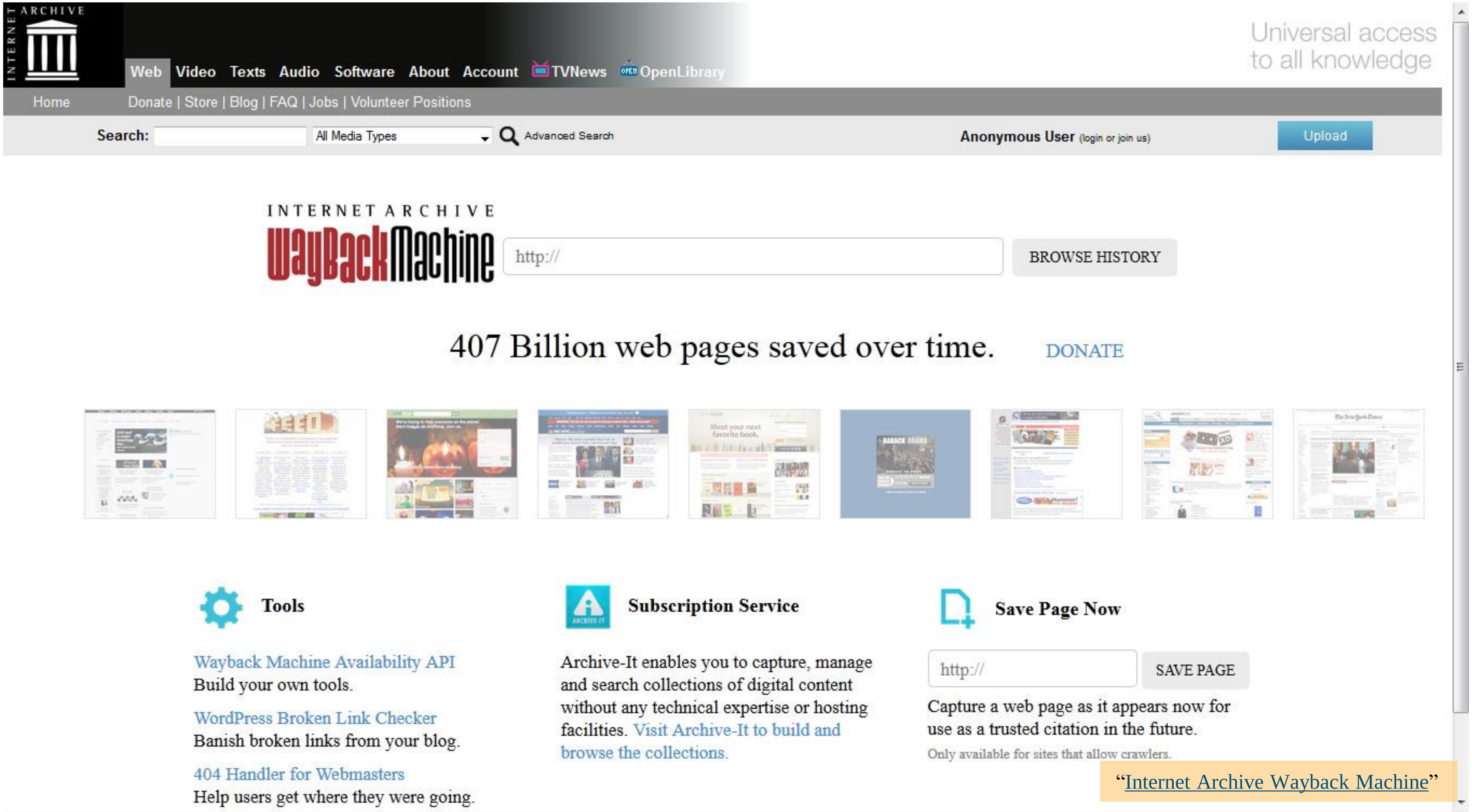
COLLECTION EFFORTS

"Beehives and bees" by Carsten aus Bonn under CC BY-ND 2.0

by the numbers

- 70 web archiving initiatives on Wikipedia
- 313 Archive-It partners
- 33 CDL WAS subscribing institutions

broad but shallow



407 Billion web pages saved over time.

Tools

Wayback Machine Availability API
Build your own tools.

WordPress Broken Link Checker
Banish broken links from your blog.

404 Handler for Webmasters
Help users get where they were going.

Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

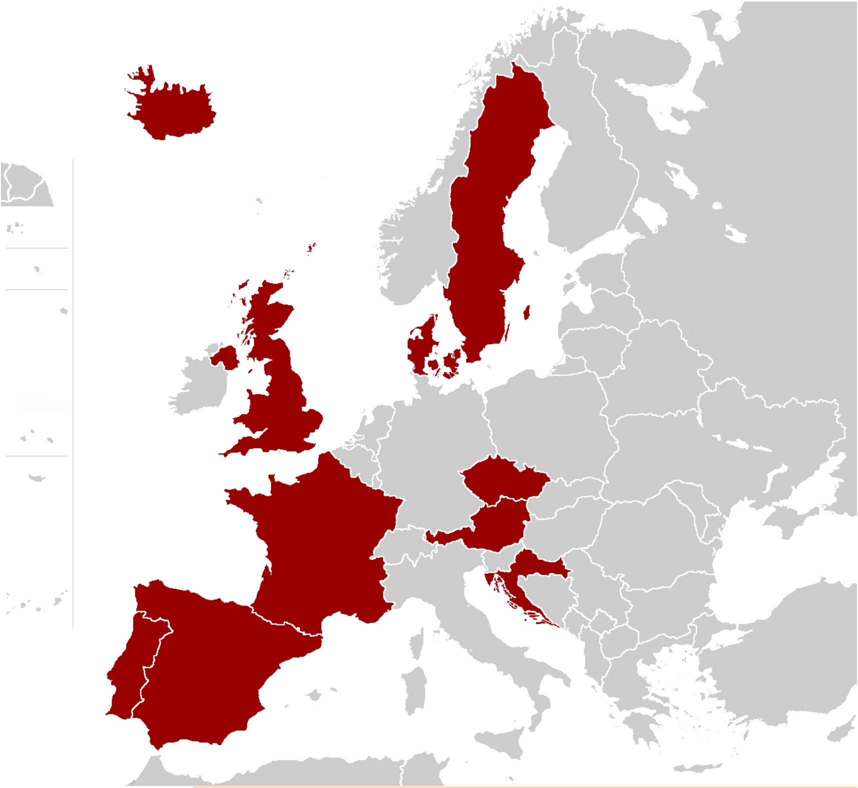
Save Page Now

Capture a web page as it appears now for use as a trusted citation in the future.

Only available for sites that allow crawlers.

"Internet Archive Wayback Machine"

national domains



"Europe blank map.png" by Wiki-vr under public domain

"AustraliaNZ-blank.png" by Chuq under CC BY-SA 3.0

Wikipedia: "List of Web archiving initiatives"

selective archiving

- Human Rights Web Archive (Columbia)
- CyberCemetery (North Texas)
- Elections Web Archives (Library of Congress)
- Labor and the Left (NYU Tamiment)
- Ukraine Conflict (Archive-It)
- NC State Government (NC Archives, Library)
- Michigan Historical Collections (UM Bentley)
- Health and Medicine Blogs (NLM)

how much archived?

dmoz

79%

bitly

16%

del.icio.us

68%

YAHOO!®

19%

"How Much of the Web Is Archived?" by Ainsworth, AlSum, SalahEldeen, Weigle, and Nelson (2011).

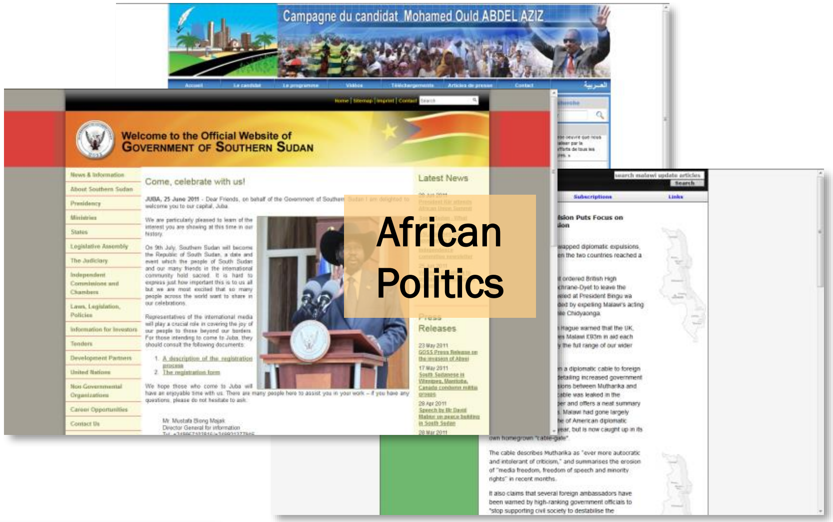
What

WE ARE COLLECTING

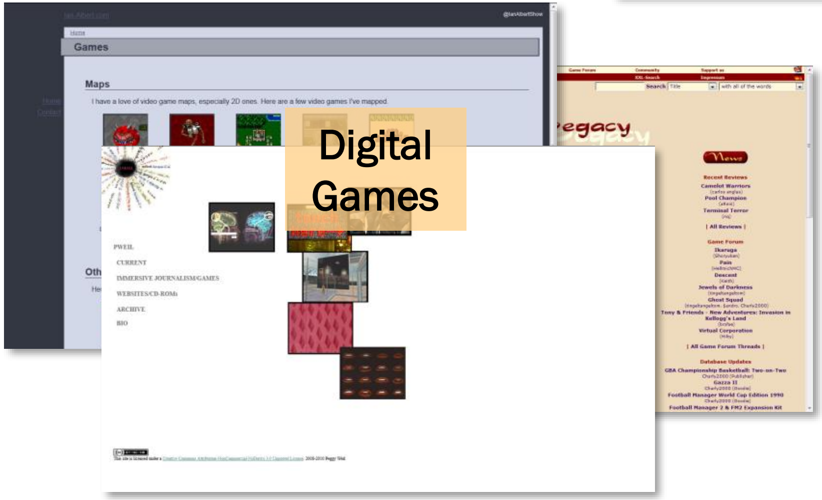
“Contents display” by orionpozo under CC BY 2.0

topical collections

Middle East Politics


Campagne du candidat Mohamed Ould ABDEL AZIZ
Welcome to the Official Website of Government of Southern Sudan
Come, celebrate with us!
Latest News
Press Releases

African Politics

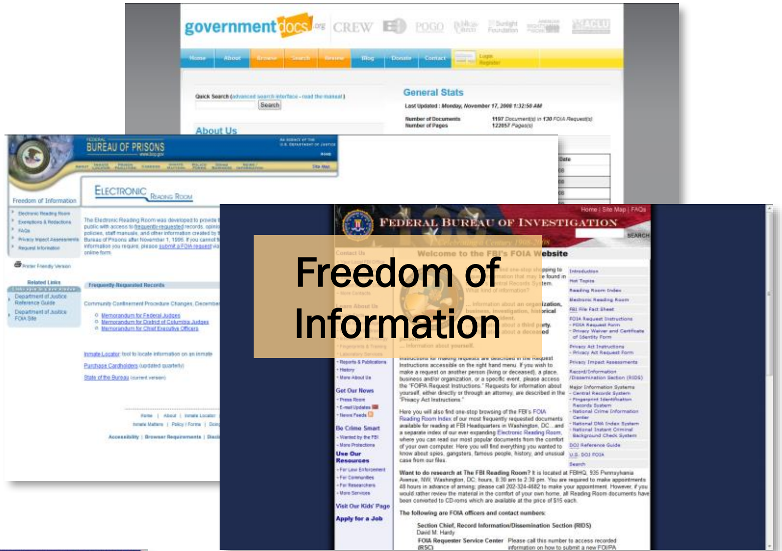


Digital Games

government information

BDCP
Bay Delta Conservation Plan
City of East Palo Alto
Association of Bay Area Governments

Bay Area Governments

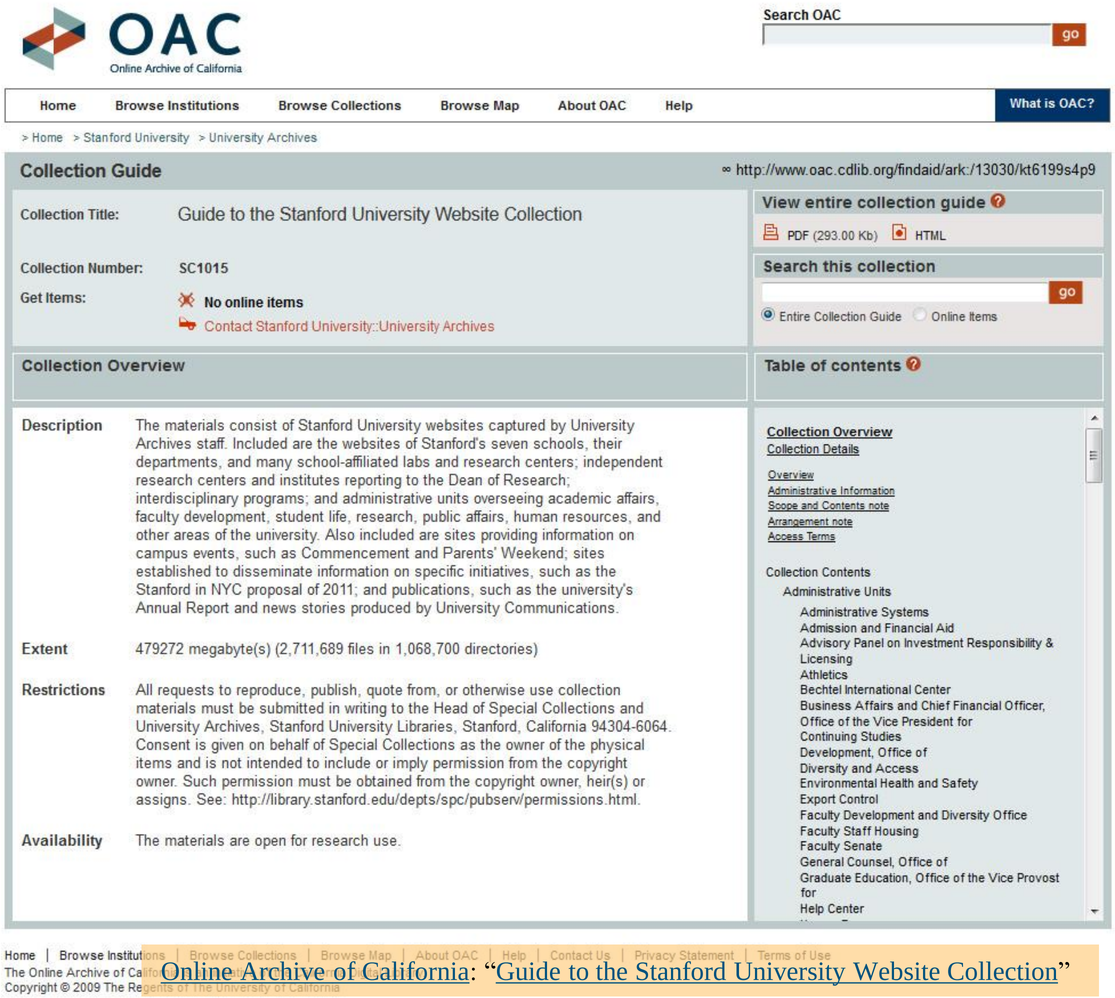

governmentdocs.org
CREW
Federal Bureau of Prisons
Electronic Reading Room
Federal Bureau of Investigation
Welcome to the FBI's FOIA Website

Freedom of Information

U.S. Department of State
By Topic
opencrs
The National Agricultural Law Center
Congressional Research Service Reports
Subject Index

CRS Reports

institutional legacy



> Home > Stanford University > University Archives

Collection Guide http://www.oac.cdlib.org/findaid/ark:/13030/kt6199s4p9

Collection Title: Guide to the Stanford University Website Collection

Collection Number: SC1015

Get Items: No online items

Contact Stanford University::University Archives

View entire collection guide

PDF (293.00 Kb) HTML

Search this collection

go

● Entire Collection Guide ○ Online Items

Collection Overview

Table of contents

Description The materials consist of Stanford University websites captured by University Archives staff. Included are the websites of Stanford's seven schools, their departments, and many school-affiliated labs and research centers; independent research centers and institutes reporting to the Dean of Research; interdisciplinary programs; and administrative units overseeing academic affairs, faculty development, student life, research, public affairs, human resources, and other areas of the university. Also included are sites providing information on campus events, such as Commencement and Parents' Weekend; sites established to disseminate information on specific initiatives, such as the Stanford in NYC proposal of 2011; and publications, such as the university's Annual Report and news stories produced by University Communications.

Extent 479272 megabyte(s) (2,711,689 files in 1,068,700 directories)

Restrictions All requests to reproduce, publish, quote from, or otherwise use collection materials must be submitted in writing to the Head of Special Collections and University Archives, Stanford University Libraries, Stanford, California 94304-6064. Consent is given on behalf of Special Collections as the owner of the physical items and is not intended to include or imply permission from the copyright owner. Such permission must be obtained from the copyright owner, heir(s) or assigns. See: http://library.stanford.edu/depts/spc/pubserv/permissions.html.

Availability The materials are open for research use.

Home | Browse Institutions | Browse Collections | Browse Map | About OAC | Help | Contact Us | Privacy Statement | Terms of Use

The Online Archive of California is an initiative of the California Digital Library

Online Archive of California: "Guide to the Stanford University Website Collection"

How OTHERS COLLECT

“2009 san diego comic-con: comics, still an elemental part of the con” by george ruiz under CC BY 2.0

necessary but not sufficient

THE TAMIMENT LIBRARY &
ROBERT F. WAGNER LABOR ARCHIVES

"In principle, the collection development policy for the Tamiment Library's Web Archive parallels that of the Tamiment Library as a whole (labor and radicalism)"

In practice, this is complicated by (a) the enormous size and variety of born digital materials within Tamiment's collecting scope…and (c) resource restraints. Thus the Library will not only have to carefully appraise materials, but to set priorities and limitations."

Tamiment Library: "Web Archiving Collecting Policy"

necessary but not sufficient

- align with organizational mission
- support research and teaching
- preserve institutional legacy
- consider history and geography

sufficient-y

- collect within subject area
- focus on at-risk content
- collect content previously collected in print
- limit to particular types of organizations

sufficient?

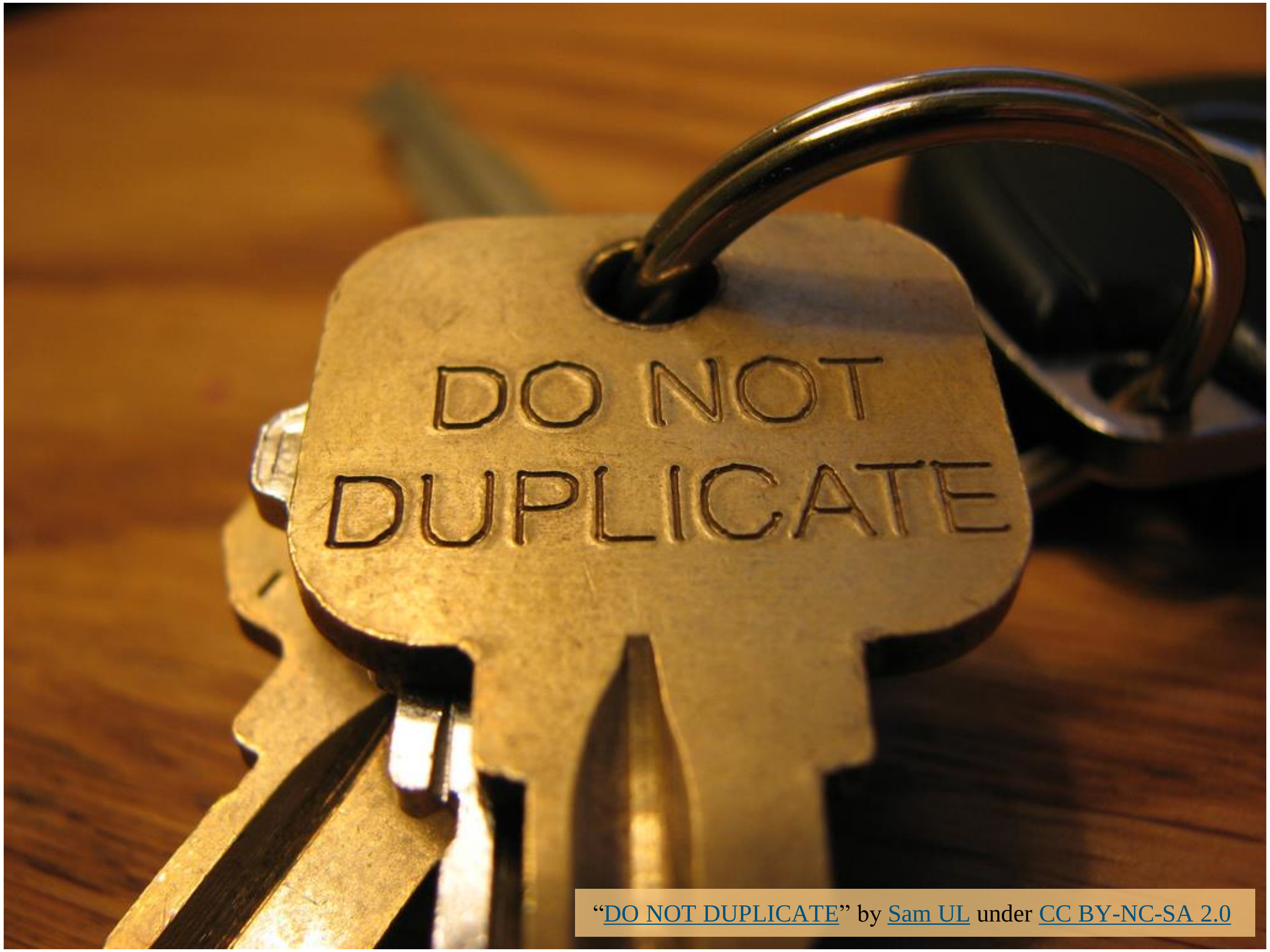
- consider what others are collecting
- don't aim to be comprehensive (if you can’t be)
- complement existing strengths
- prefer current and/or unique content
- mind resource constraints
- anticipate value to researchers
- collect content, not links to content
- only collect publicly available content
- only target specific resource or format types
- enable designated research

Additional

COLLECTION CONSIDERATIONS

“Fishing” by Wisconsin Department of Natural Resources under CC BY-ND 2.0

copyright and access policy



"DO NOT DUPLICATE" by Sam UL under CC BY-NC-SA 2.0

FERPA

"Media Center" by UW Libraries under CC BY-NC-ND 2.0

social media

AT THE SAME TIME.
AMANDA.
a better understanding of who our customers are, and what they want from Dell!
- Caroline Dug
HELP PEOPLE REALIZE THEIR VOICE AND IDENTITY
KENNY LAUER
George P. Johnson
ENTRAR EM CONTATO COM CLIENTES AO REDOR DO MUNDO E ME TORNAR UMA CIDADÃ GLOBAL!
ANA EVANGELISTA
DELL BRASIL
... Listen, learn and Engage!
The Pulse of the Customer

“Social Media & Community gives me the power to...” by Dell Inc. under BY-NC-ND 2.0

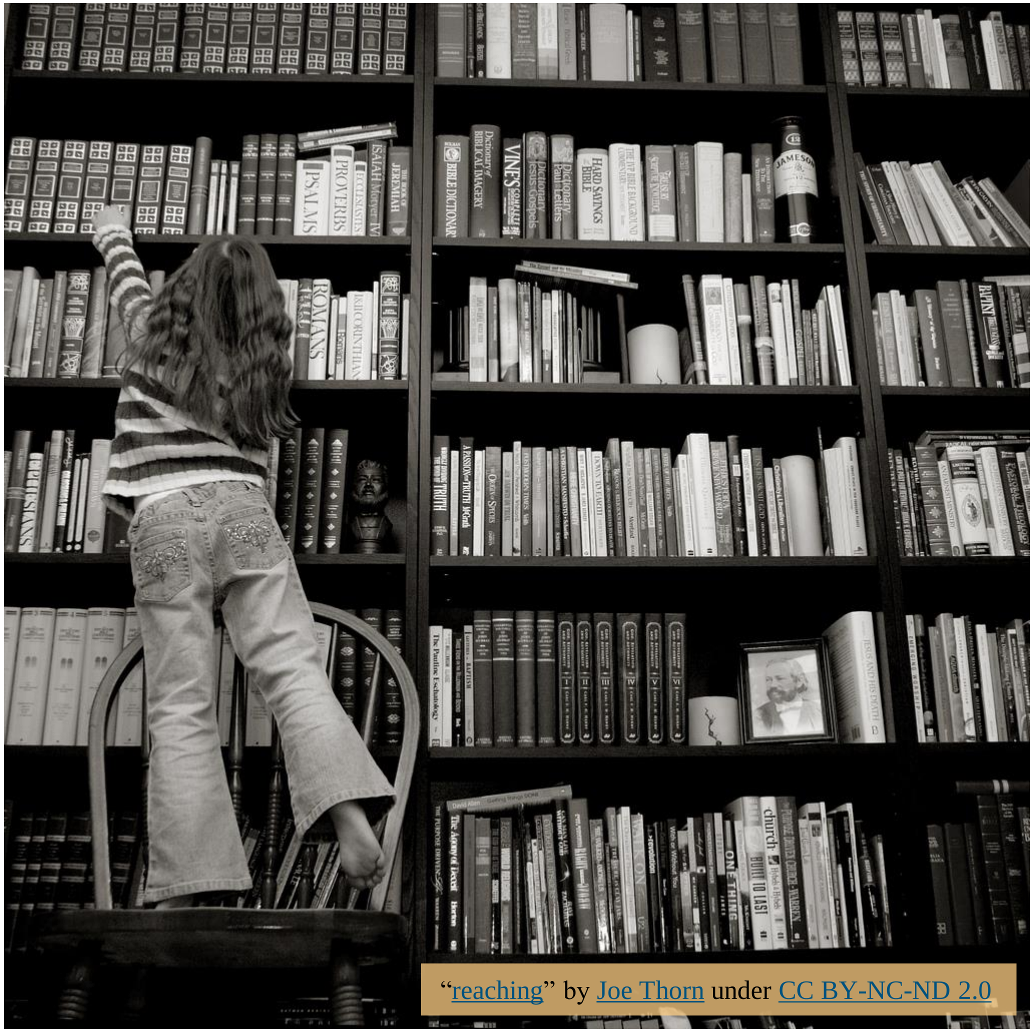
archiving proscriptions

"Downloading is Prohibited!" by Oliver Hine under BY-NC-ND 2.0

offramps

"Pathfinder Panorama" by NASA Solar System Exploration under Public Domain

technical challenges



“reaching” by Joe Thorn under CC BY-NC-ND 2.0

cost modeling

“dollar butterfly (2)” by eikosi under BY-SA 2.0

collection use cases

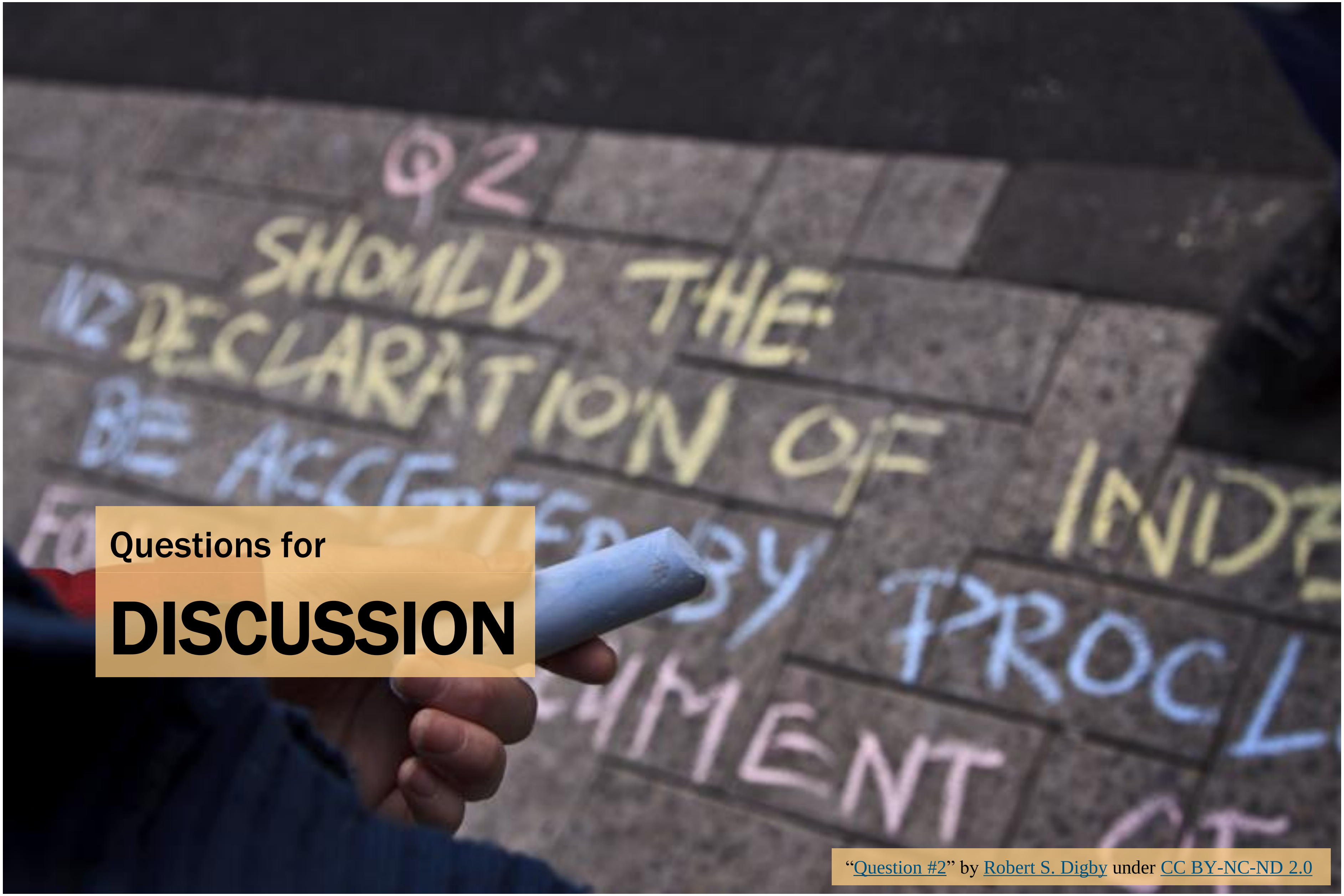
- outreach and education
- persistent citation
- documenting spontaneous events
- preserving citizen journalism
- saving at-risk content
- litigation risk mitigation
- capture related resources
- records management

"Web Archiving Use Cases" by Emily Reynolds (2013)

research use cases

- how incumbent candidates talk to local constituencies
- inter-link graph of websites in different languages
- policies and practices of public health NGOs
- Honduran government websites after 2006 coup
- file format analysis for preservation planning
- prevalence of semantic markup
- most commonly-used JavaScript libraries
- digital archaeology of GeoCities
- resource persistence in Egypt Revolution social media

"Web Archiving Use Cases" by Emily Reynolds (2013)



Questions for

DISCUSSION

“Question #2” by Robert S. Digby under CC BY-NC-ND 2.0

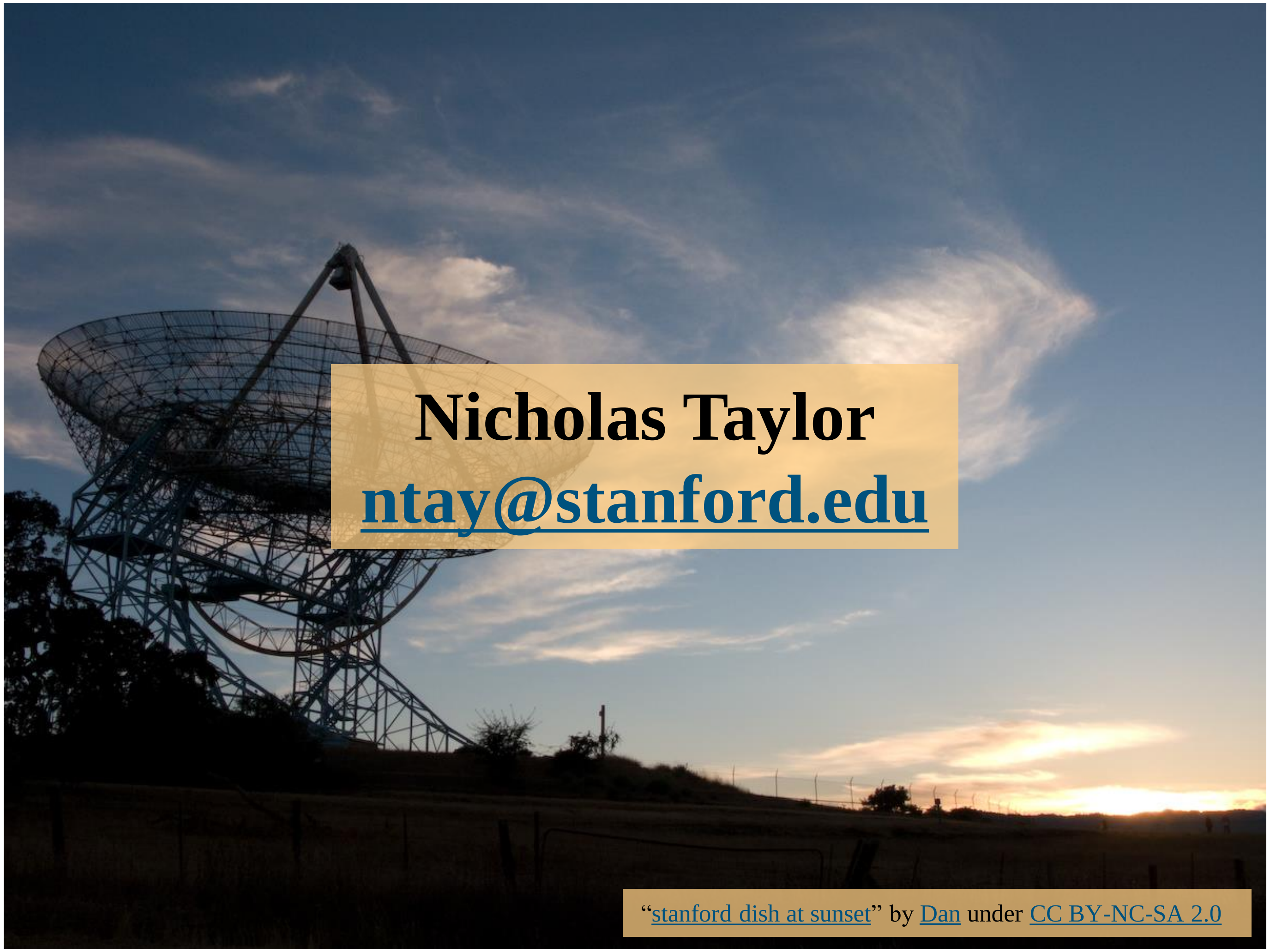
where do we start?

1. maintain awareness of topical web archives
2. promote and facilitate access to existing web archives
3. provide curatorial assistance for collaborative projects
4. enhance our collection dev policies for web archives
5. evaluate local/global gaps to build unique collections

how do we do it?

- how can we maintain awareness, facilitate discovery, and promote use of other web archives?
- what is relative importance of collection development policies, at-risk nature of the content, research use case tangibility, what others are collecting, etc.?
- how to maximize value of existing web archives and those we create, for Stanford and larger community?
- what are the key elements and optimal approach for creating our own collection development policies?

thank you!



Nicholas Taylor

ntay@stanford.edu

“stanford dish at sunset” by Dan under CC BY-NC-SA 2.0