STATE OF CALIFORNIA WEBSITE TRENDS 2008-2010

Tracy Seneca
California Digital Library
tracy.seneca@ucop.edu
April, 2010

TABLE OF CONTENTS

FIGURES, TABLES AND GRAPHS

OVERVIEW

In October 2008 the California Digital Library (CDL) began capturing the agency and department Web sites of the State of California using CDL's Web Archiving Service. The list of sites was derived from an online directory of State agencies and after duplicated sites (variations on agency names) were removed, there were 304 State government sites included in the archive. [1] This number changes as new Web sites emerge and as existing sites disappear. Captures were run of the full archive in October 2008, April 2009 and February 2010. Public access to the **California State Government : .ca.gov Web Archive** is available at: http://webarchives.cdlib.org/calgov . In deference to copyright recommendations, the public archive provides access to Web content that is at least six months old, and observes an embargo on the most recently captured materials.

The goal of this work is both ambitious and achievable; to provide archival access to the Web presence of the State of California for researchers at the University of California, for California citizens, and for the State agencies themselves. California is not only the world's 8th largest economy,[2] it is also the U.S. government's 3rd largest Web domain[3], and in preserving it for future researchers, CDL and the University of California are providing an extraordinary service to the State of California.

This archive also supports ongoing collaborative work among the University of California Libraries. Beginning in July 2008, librarians at the University of California campuses began to assess the feasibility of a collaborative approach to identifying and cataloging key born-digital State of California publications.[4] As part of this project, they identified a watch-list of 33 California State agency sites known to frequently publish materials critical to UC research and scholarship. CDL consequently ran more frequent captures of the sites the librarians identified as being most critical. CDL staff also worked with catalogers in the project to explore cataloging workflows that point to archived versions of the documents, to insure that the documents remain available via the catalog even if they are removed from the live website. While the archive provides stable targets for these catalog records, the cataloging activity also improves the archive. If for some reason the archive does not contain the State publication in question, the cataloger can add it.

Owing to this work, as of the spring of 2010 we now have a large body of data that can tell us a good deal about State of California website trends over the last year and a half. With several hundred captures in the archive, each one including reports such as site size and file types captured, we can begin to answer questions such as:

- Which California agency sites are actually the largest? (By storage size? By number of files?)
- Which sites offer the most PDF files (assuming these correspond to meaningful publications)?
- Which sites grew most dramatically in this time frame?
- Which sites lost the most content (all file types or just PDF files) in this time frame?
- Which agencies prevent their sites from being captured effectively (or entirely) by an archive?

The following report presents the results of our archive analysis in two distinct sections. Section 1: "California Agency Web Server Rules and the Archive" will demonstrate the profound impact that State agency server settings (usually determined by a system administrator) have on the ability to archive State content. This is a more qualitative assessment of the materials in the archive.

[1] http://www.ca.gov/About/Government/agencyindex.html
[2] "Facts on Policy: The California Economy". Hoover Institution, Stanford University. http://www.hoover.org/research/factsonpolicy/facts/35975159.html
[3] Shraddha Ladda and Subhashri Suresh. "Weblab - .gov domain analysis". May 2009. http://weblab.infosci.cornell.edu/papers/Ladda2009.pdf
[4] This is in conjunction with the work of the UC Shared Cataloging Project http://www.cdlib.org/services/collections/scp/

Section 2: "Archive Analysis" is a more quantitative analysis and examines overall site change, with particular emphasis on the 33 priority agency sites identified by UC librarians. The overall goal of this report is not just to provide a snapshot of the California State Government Web presence, but also to demonstrate just some of the things that can be learned about a Web domain once it has been archived.

SECTION 1: CALIFORNIA AGENCY WEB SERVER RULES AND THE ARCHIVE

One aspect of Web site management that profoundly affects Web archiving is a server standard called *robots.txt exclusion files*[5]. These are files that a site owner can place on a web server that provide instructions for Web crawlers. When visiting a site, a "well-behaved" Web crawler will first look for a file named [site homepage URL]/robots.txt. If a file is found, and if it contains instructions about capturing the site, the "well-behaved" crawler will observe those rules.

Site owners have a good deal of control over what they can specify in robots.txt files. They can decide to limit access to particular files, they can prohibit entire directories from capture, or they can prevent crawlers from capturing an entire site. They can also selectively apply rules to particular crawlers. Here is what it looks like to a curator when a site is entirely prevented from capture by a robots.txt file.

FIGURE 1: CALIFORNIA STATE CONTROLLER'S OFFICE

On April 10 2009, the WAS crawler archived 9,663 documents from the California State Controller's site. On June 12th 2009, it captured only three. These three files include 2 files needed to locate the site, plus an archival copy of

[5] http://www.robotstxt.org/

the robots.txt file itself, containing the rules to crawlers. Those rules are included above and state that Google and one other Web crawler may capture the site, but no one else may do so.

The California Digital Library currently configures the Web Archiving Service crawler to respect these rules. In their spring 2008 recommendations, the Section 108 Study Group[6] asserted that government agencies and political candidates should *not* be able to prohibit libraries from capturing their Web content. While CDL endorses this position, it is still only a recommendation. There are other, less transparent ways that a site owner can prohibit access to Web crawlers, and until we have a more clear exception under Section 108 of the Copyright law, we are bound to observe the robots.txt exclusion rules.

The consequences of this are enormous. The University of California is archiving the Web publications of the State of California in the midst of the worst fiscal crisis the State has known. The California State Controller's Office illustrated above is one of the agencies standing at the heart of that crisis, and as of June 2009, because of the robots.txt restrictions added to the site, we are no longer able to archive that material for scholarly research.

In September 2009, CDL analyzed the California State Agency archive to assess the impact of robots.txt files. For every site, we evaluated the robots.txt file (if found) and assessed the quality of the archived content.

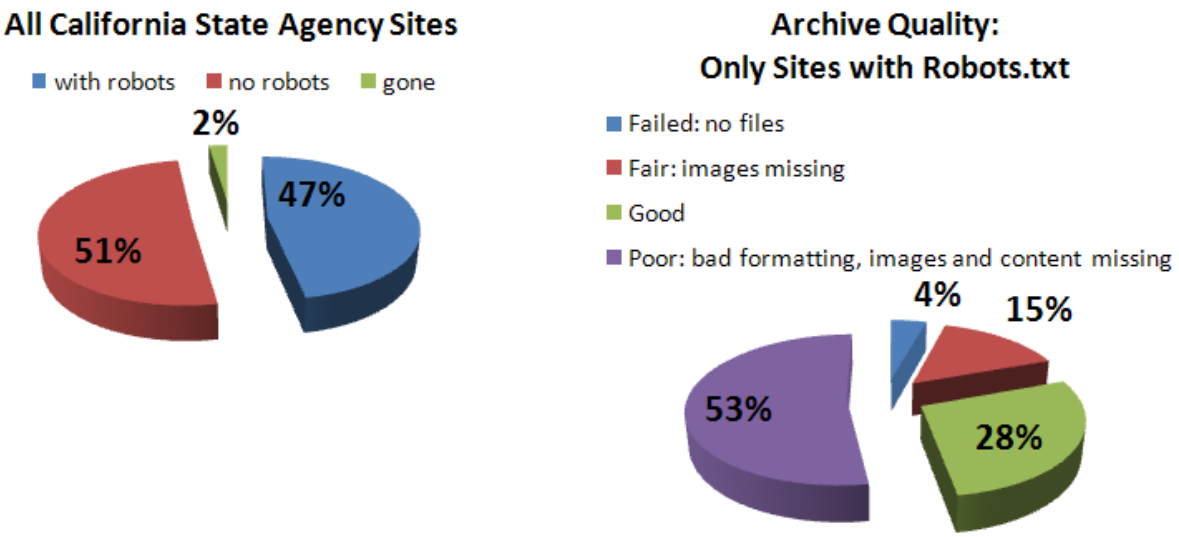

FIGURE 2: IMPACT OF ROBOTS.TXT RULES ON CALIFORNIA AGENCY ARCHIVE

Overall, 47% of California agency sites contained a robots.txt file with Web crawler rules and in 4% of those cases, they entirely prevented the archiving of the site. In themselves, robots.txt files may not adversely impact the quality of an archive. In some cases these files may be expressly written to prevent access to test areas of the server. In 28% of the cases where a robots.txt file appeared, the archived version of the site was still high quality and contained the site's substantive material.

As mentioned, in addition to preventing sites entirely from capture, agencies can prevent the capture of specific files or directories, resulting in an archived version that is either unusable or devoid of substantive content. Robots.txt files may contain lines such as:

[6] The details of their recommendations can be found on pages 80-87 of the **Section 108 Study Group Report:** http://www.section108.gov/docs/Sec108StudyGroupReport.pdf

Disallow:/docs
Disallow:/pubs
Disallow:/images
Disallow:/styles

These rules are likely to result in an archival copy of the site that cannot be rendered or used effectively by future researchers. Rules such as these are may put in place by a system administrator who wishes to make the site highly available to search engines such as Google by making the text accessible to Web crawlers, but who fears that there will be too much server traffic if crawlers have access to files that are not "necessary" to search engines. This is a relatively standard practice which unfortunately makes it impossible to effectively archive a site.

15% of the California Agency sites with robots.txt files were missing images, and another 53% were missing critical data needed to format the site for viewing, resulting in an archived version that is hard to navigate and is not true to the original website. Of the sites with poor archival quality, 28 of them had identical robots.txt files, even though the agencies were not immediately related:

User-agent: *
Disallow: /images
Disallow: /classes
Disallow: /cgi-bin
Disallow: /htdig
Disallow: /js
Disallow: /styles
Disallow: /ssi
Disallow: /css
Disallow: /javascript

This would appear to be a template file, provided by an office that supports the delivery of Web content for State agencies. That default template, however, excludes any of those agencies' sites from being effectively archived for future researchers. There were no obvious trends in the nature of an agency versus how open or restrictive the exclusion rules were.

Not restricted:

- Office of Information Security and Privacy Protection
- Office of Systems Integration
- Legislative Analyst's Office

Restricted:

- Reporting Transparency in Government
- California State Library
- Office of State Publishing
- Legislative Counsel

Any state agency that would like to expressly *permit* the California Digital Library to archive their website should add the following two lines to the top of their robots.txt file:

```
User-Agent: cdlwas_bot
Disallow:
```

SECTION 2: ARCHIVE ANALYSIS

In the tables and charts below, we compare the April 2009 and February 2010 captures of the entire list of about 300 State agency sites. Websites were captured in these two months using exactly the same crawler settings, so any change between them is more likely to reflect actual change on the site. In both cases, we had the crawlers capture each site as completely as possible, without capturing content from external sites that might a site might link out to.

In the tables that follow, if an agency name is highlighted, that means it is one of the 33 priority agencies identified by UC librarians. The word "files" means any distinct file captured, whether an individual image, an HTML page or PDF document.

There are a couple of additional things to note about site statistics before looking at the data. A "loss" in size, files or PDF files means that fewer materials were captured in 2010 than in 2009. The 2009 files, of course, remain in the archive. Further, the decrease does not always mean that files were removed from the live website. If the site administrator altered the server rules to restrict crawler access to specific areas, this would result in fewer documents archived, but the documents may still remain on the live Web. Likewise, a dramatic growth in the number of files does not always mean that the website grew in size, particularly if the growth is strictly in HTML files. There are some website design practices that can create "crawler traps" of relatively meaningless data – such as calendars that can be navigated endlessly into the future. Whenever an extreme gain or loss in site statistics occurs, this is a cue to look more closely at the archived content and determine if the change is meaningful.

Finally, the statistics concerning gain and loss in this analysis are simply overall number counts; they do not account for content that may have changed dramatically from one date to the next. If the content of a site has been dramatically altered, but did not result in a noticeable change in the number of files it provides, it would not stand out in these figures. The Web Archiving Service does offer tools for examining more substantive change on a file by file basis for each site, but this analysis is intended to provide more of a birds-eye view of the entire archive.

THE PRIORITY LIST OF AGENCIES

Here is the list of priority agency sites identified for the State of California:

Agency	Campus	Files 2009	Files 2010	PDFs 2009	PDFs 2010	change
Department of Water Resources	UC Berkeley	48030	13112	3821	4839	-34918
Department of Toxic Substances Control	UCLA	47159	24775	7103	7567	-22384
California Emergency Management Agency (was Governor's Office of Emergency Services)	UCLA	26741	6848	2218	1413	-19893
California Integrated Waste Management Board	UC San Diego	80656	65956	3411	4716	-14700
State Water Resources Control Board	UC Berkeley	49251	36688	34431	29106	-12563
California State Controller	UC San Diego	9863	3	8317	0	-9860
California Energy Commission	UC Berkeley	19641	14316	9661	6572	-5325
California Department of Education	UC Davis	3016	7	271	0	-3009
Postsecondary Education Commission	UC San Diego	10274	7308	1276	1002	-2966
Board of Equalization	UC San Diego	22192	20529	18577	16560	-1663
California Policy Research Center	UC Berkeley	138	0	69	0	-138
Senate Office of Research	UC Davis	3	3	0	0	0
Franchise Tax Board	UC Berkeley	14354	14367	9273	9286	13
Division of communicable disease control	UC Irvine	227	242	0	0	15
Governor's Office of Planning and Research	UCLA	512	560	431	476	48
Little Hoover Commission	UC Berkeley	1857	2058	1013	1046	201
California Coastal Commission	UC Berkeley	3662	3880	316	337	218
Legislative Analyst's Office	UC Davis	3283	3648	0	0	365
Department of Public Health	UC Irvine	243	614	0	0	371
Department of Pesticide Regulation	UCLA	13881	14319	7763	8284	438
Office of Environmental Health Hazard Assessment	UC Davis	5371	5816	2781	3133	445
California Volunteers (was Governor's Office of Service and Volunteerism)	UCLA	6069	6517	62	111	448
Department of Finance	UC Irvine	7309	7833	2085	2394	524
Secretary of State	UC Irvine	10228	10850	5934	7102	622
Division of Occupational Safety & Health (Cal/OSHA)	UC Irvine	3804	4794	1034	1310	990
Division of Labor Statistics and Research	UC Irvine	14063	15326	7735	8536	1263
CALFED Bay-Delta Program	UC Davis	7689	9863	2358	2384	2174
Legislative Counsel	UCLA	126	2487	0	0	2361
Bureau of State Audits	UC San Diego	5581	9691	2242	2366	4110
Labor Market Information Division	UCLA	113176	133190	54	1106	20014
California Department of Food and Agriculture	UC Davis	11762	51323	0	0	39561
California Air Resources Board	UC San Diego	52711	132831	30580	28803	80120

FIGURE 3: PRIORITY AGENCY SITE STATISICS

To determine if the agency sites are evenly distributed among the UC campuses, a total of the 2010 # of files for 2010 shows a distribution of:

Campus	2010 Files
UC Irvine	39,659
UC Davis	70,660
UC Berkeley	84,421
UCLA	188,696
UC San Diego	236,318

FIGURE 4: DISTRIBUTION OF WORK BY CAMPUS

Note that the "Demographic Research Unit" was also listed as a priority site to be monitored by UC Irvine, but it is a directory within the Department of Finance, also a priority site for Irvine. The Demographic Research Unit content is automatically included with the Department of Finance site.

The table shown in figure 1 is organized in order by the degree to which the sites have changed from 2009 to 2010. Sites at the top of the list lost the most documents, sites at the bottom gained the most documents, and the sites in the middle are the priority agencies that didn't show a great deal of change.

The impact of site owner server rules on the collection of these priority sites is examined separately in the *California Web Agency Server Rules and the Archive* section of this report, starting on page 19.

ROBOTS.TXT FILES AND THE PRIORITY LIST OF AGENCIES

In February 2010, a closer examination of the impact of Robots.txt files was conducted against the archived versions of the priority agency list for the University of California Librarians. 65% of the priority sites contained robots.txt files. Of those, 43% were damaging to the quality of the archived site.

Among the more interesting restrictions found:

- The 2000 Edition of the Guide to School Site Analysis and Development may not be archived. http://www.cde.ca.gov/ls/fa/sf/documents/schoolsiteanalysis2000.pdf may not be captured by a crawler from the California Department of Education. This is the CDE site's only restriction. The document is, however, freely available on the Web via ERIC.
- The Senate Office of Research site expressly allows crawling by the Internet Archive, but prohibits all others.
- The California Department of Water Resources only prohibits the capture of the data sets and publications available via the water data library (http://www.water.ca.gov/waterdatalibrary/docs/Hydstra/). These would arguably be the most critical resources to preserve in a Web archive.

Agency	Robots.txt	Campus
State Water Resources Control Board	fair	UC Berkeley
California Integrated Waste Management Board	poor	UC San Diego
California State Controller	failed	UC San Diego
California Energy Commission	poor	UC Berkeley
Senate Office of Research	failed	UC Davis
Franchise Tax Board	poor	UC Berkeley
Governor's Office of Planning and Research	poor	UCLA
Department of Pesticide Regulation	poor	UCLA
Secretary of State	failed	UC Irvine
Division of Occupational Safety & Health (Cal/OSHA)	poor	UC Irvine
Division of Labor Statistics and Research	poor	UC Irvine
CALFED Bay-Delta Program	poor	UC Davis
Legislative Counsel	poor	UCLA
Department of Water Resources	good	UC Berkeley
California Department of Education	good	UC Davis
Office of Environmental Health Hazard Assessment	good	UC Davis
California Volunteers (was Governor's Office of Service and Volunteerism)	good	UCLA
Bureau of State Audits	good	UC San Diego
Labor Market Information Division	good	UCLA
California Air Resources Board	good	UC San Diego
California Policy Research Center	site removed	UC Berkeley
Department of Toxic Substances Control	none	UCLA
California Emergency Management Agency (was Governor's Office of Emergency Services)	none	UCLA
Postsecondary Education Commission	none	UC San Diego
Board of Equalization	none	UC San Diego
Division of communicable disease control	none	UC Irvine
Little Hoover Commission	none	UC Berkeley
California Coastal Commission	none	UC Berkeley
Legislative Analyst's Office	none	UC Davis
Department of Public Health	none	UC Irvine
Department of Finance	none	UC Irvine
California Department of Food and Agriculture	none	UC Davis

FIGURE 5: IMPACT OF ROBOTS.TXT RULES ON THE PRIORITY CALIFORNIA STATE AGENCY SITES

failed = no files
poor = bad formatting, images or content missing
fair = some content missing but useable

Sites Lost or Gained in Entirety

The following table shows 11 State agency sites captured in 2009 that could no longer be captured at the same address in 2010. In cases where the number of 2010 documents = 0, the home page URL for the site had simply stopped working. In cases where the number of 2010 documents = 3, the server administrator had posted rules to prevent the site from being captured. (This list does not include the site that had been prevented from capture all along). Of the sites with home page URLs that no longer returned a result, only two appear to have simply disappeared from the Web, however one of those was a priority agency.

Agency	# of Files		Notes
	2009	2010	
California Policy Research Center	138	0	Cannot find evidence of site or formal closure of service
Telecommunications Division	70	0	Reorganized, now the Public Safety Communications Division w/in OCIO. Captured as part of the OCIO site.
Crime and Violence Prevention Center	2434	0	Division shut down. No forwarding of services.
Division of Land and Right of Way	27	0	Reorganized; now captured as part of the Division of Engineering site.
Committee on Dental Auxiliaries	308	0	Abolished on July 1 2009, regulation now occurs under Dental Board of California, captured separately.
CalGOLD (Business Permit Information)	77	3	
California State Controller	9863	3	
Coastal Conservancy	3543	3	
Office of Lieutenant Governor	2498	3	
California Workforce Investment Board	52016	3	
Electricity Oversight Board	94	3	

FIGURE 6: SITES ARCHIVED IN 2009 NO LONGER ACCESSIBLE IN 2010

Comparison by Number of Files Captured 2009 & 2010

The following tables and graphs rank the top thirty sites by the number of files for 2009 and 2010. The graphs illustrating the top 30 sites for April 2009 and Feb 2010 will show how much each site has changed in size, so you can easily see which sites grew or shrank. Figure 7 shows a table ranking the 30 state agencies with the largest percentage of growth in documents; this will surface the smaller sites that still showed significant growth. Figure 8 lists the sites that lost the largest percentage of files, and will demonstrate even the smaller sites that may have shrunk significantly. Again, the discrepancy in size may be due to new restrictions on the agency server or a website design change that makes capture more difficult; it may not reflect actual document loss on the Web. That said, figures such as these in generated from an archive will at least provide the curator with a good cue for sites to investigate.

2009		2010	
CA Agency	# files	CA Agency	# files
Labor Market Information Division	113176	Labor Market Information Division	133190
Department of Industrial Relations	91337	California Air Resources Board	132831
Assembly Democratic Caucus	88669	Assembly Democratic Caucus	132042
California Integrated Waste Management Board	80656	Department of Industrial Relations	86287
Governor's Office	75836	CalPERS	70492
CalPERS	73776	California Integrated Waste Management Board	65956
State Treasurer	58782	Office of the Inspector General	59344
Office of Technology Services	58266	State Treasurer	58155
State Bar of California	56076	State Bar of California	55116
Office of the Inspector General	55625	California State Science Fair	54884
California Air Resources Board	52711	California Department of Food and Agriculture	51323
California Workforce Investment Board	52016	Governor's Office	46518
State Water Resources Control Board	49251	Regional Water Quality Control Boards	42709
Regional Water Quality Control Boards	48610	Central Valley Bay-Delta Branch	40500
Department of Water Resources	48030	Office of Technology Services	39980
Department of Toxic Substances Control	47159	CAL FIRE	39859
Department of Transportation (Caltrans)	34432	Department of Fish and Game	37579
California Courts	30707	California Data Exchange Center	36711
CAL FIRE	29267	State Water Resources Control Board	36688
California Emergency Management Agency	26741	Department of Transportation (Caltrans)	36612
Standardized Testing and Reporting (STAR)	26690	State Park and Recreation Commission	33099
Department of Fish and Game	24042	California State Parks	32627
California Community Colleges System Office	23094	California Courts	30374
Board of Equalization	22192	California Office of Historic Preservation	29620
California State Science Fair	20913	California Community Colleges System Office	28618
California Data Exchange Center	20134	Standardized Testing and Reporting (STAR)	26619
California Energy Commission	19641	Fair Political Practices Commission	26269
California Maritime Academy	15131	California Institute for Regenerative Medicine	26207
Franchise Tax Board	14354	Department of Toxic Substances Control	24775
Office of Statewide Health Planning and Dev	14242	Museum Resource Center	24248

FIGURE 7: COMPARISON OF 2009 AND 2010 TOP 30 CA AGENCIES BY NUMBER OF FILES

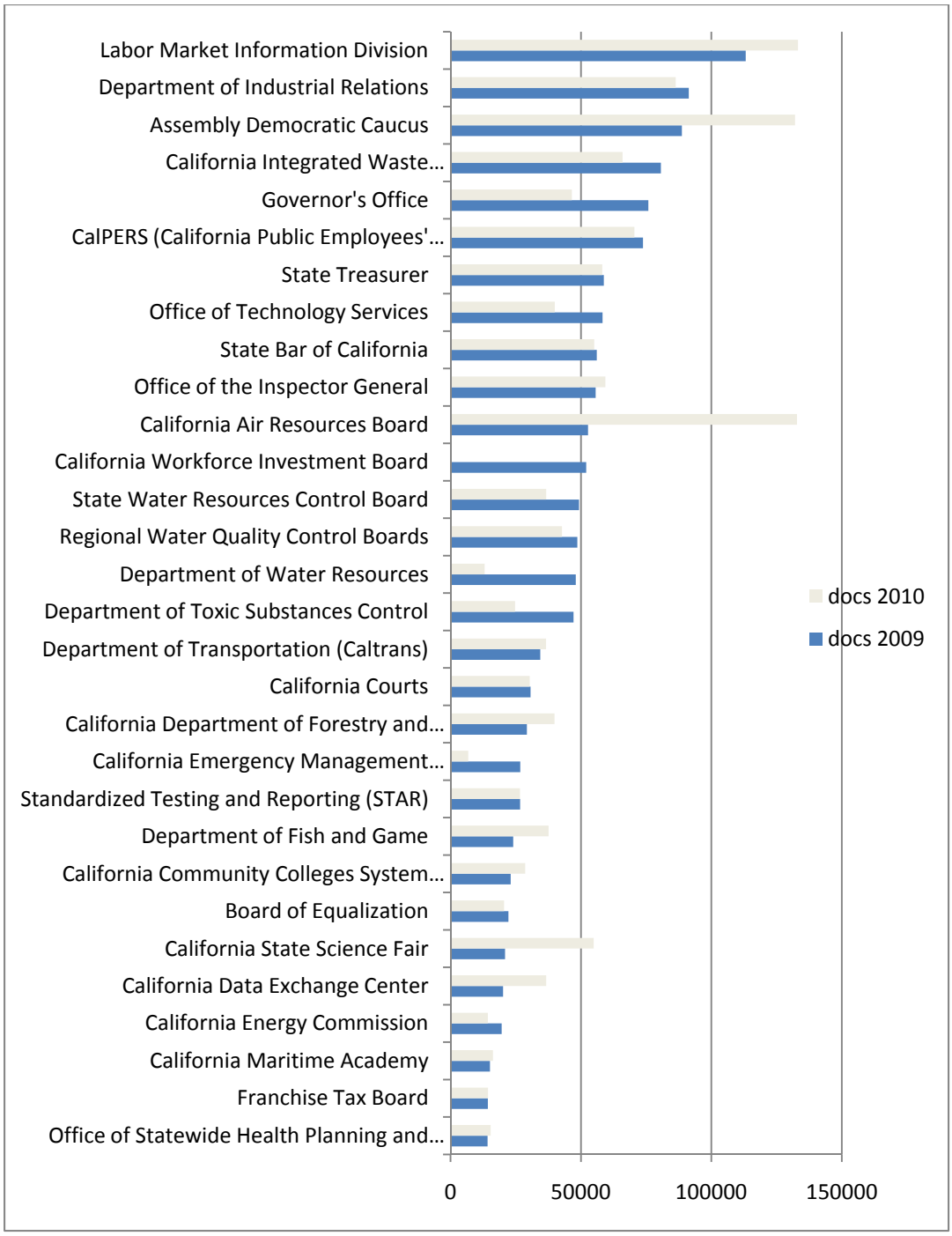

FIGURE 8: LARGEST CA AGENCY SITES BY # OF FILES (DOCS), APRIL 2009

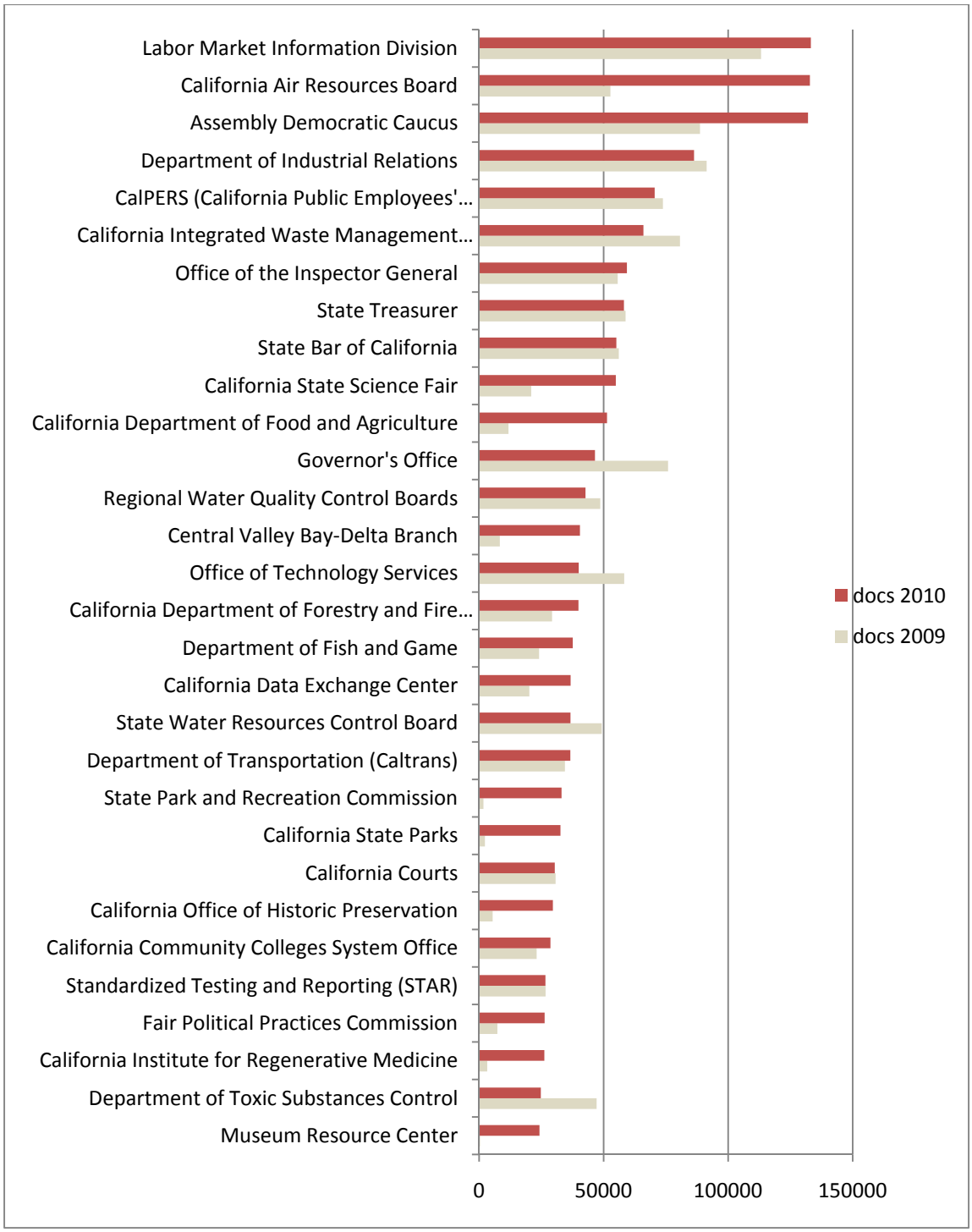

FIGURE 9: LARGEST CA AGENCY SITES BY # OF FILES (DOCS), JANUARY 2010

agency	Files 2009	Files 2010	% growth
Reporting Transparency in Government Website	0	4089	100
Public Safety Communications Division	0	340	100
California Economic Recovery Portal	0	619	100
California State Fair	3	1052	99.7
Museum Resource Center	376	24248	98.4
Cal Atlas: Geospatial Clearinghouse	154	6942	97.8
Board of Forestry and Fire Protection	16	458	96.5
Legislative Counsel	126	2487	94.9
State Park and Recreation Commission	1749	33099	94.7
California African-American Museum	794	12404	93.6
California State Parks	2337	32627	92.8
High-Speed Rail Authority	272	2708	90
California Institute for Regenerative Medicine	3234	26207	87.7
California Office of Historic Preservation	5399	29620	81.8
Central Valley Bay-Delta Branch	8286	40500	79.5
California Department of Food and Agriculture	11762	51323	77.1
Fair Political Practices Commission	7301	26269	72.2
CA Infrastructure and Economic Development Bank	160	548	70.8
State Chief Information Officer (CIO)	2399	7420	67.7
California State Science Fair	20913	54884	61.9
Department of Public Health	243	613	60.4
California Air Resources Board	52711	132831	60.3
State and Consumer Services Agency	345	765	54.9
Assembly Republican Caucus	4277	9272	53.9
Division of Recycling	268	573	53.2
California Conservation Corps	1156	2375	51.3
California Department of Veterans Affairs	1486	2930	49.3
California Board of Occupational Therapy	232	454	48.9
California Science Center	2900	5640	48.6
California Women's Commission	9648	18030	46.5

FIGURE 10: HIGHEST PERCENTAGE OF GROWTH, NUMBER OF DOCUMENTS 2009-2010

agency	Files 2009	Files 2010	% loss
California State Controller	9863	3	100
Committee on Dental Auxiliaries	308	0	100
California Policy Research Center	138	0	100
Crime and Violence Prevention Center	2434	0	100
Division of Land and Right of Way	27	0	100
OTAN (Outreach and Technical Assistance Network)	3996	0	100
California Workforce Investment Board	52016	3	100
Telecommunications Division	70	0	100
Office of Lieutenant Governor	2498	3	99.9
California Department of Education	3016	7	99.8
Coastal Conservancy	3543	6	99.8
Operations Control Office	1202	24	98
California Commission for Jobs and Economic Growth	166	4	97.6
Electricity Oversight Board	94	3	96.8
CalGOLD (Business Permit Information)	77	3	96.1
Office of Information Security and Privacy Protection	767	42	94.5
Santa Monica Mountains Conservancy	2508	571	77.2
California Emergency Management Agency	26741	6848	74.4
Department of Water Resources	48030	13112	72.7
California State Railroad Museum	6587	1967	70.1
Career Resource Network	402	131	67.4
California Environmental Protection Agency (Cal/EPA)	9100	3474	61.8
Office of Fleet Administration	143	57	60.1
E-Procurement: California State Contracts Register	131	54	58.8
Department of Toxic Substances Control	47159	24775	47.5
California State Summer School for Mathematics and Science (COSMOS)	258	138	46.5
Department of Child Support Services	6736	3727	44.7
Governor's Office	75836	46518	38.7
California Natural Resources Agency	2108	1356	35.7
Office of Technology Services	58266	39980	31.4

FIGURE 11: HIGHEST PERCENTAGE OF DOCUMENT LOSS 2009-2010

COMPARISON BY NUMBER OF PDF FILES 2009 & 2010

The following tables and graphs examine change in the number of PDF-format files. While change to a large number of HTML files might merely indicate a site redesign, change in the number of PDF files might reflect a more substantive change to the information a site provides. In some cases, however, it may simply reflect an agency's policies of access. The "Horse Racing Board" for instance, consistently provides a large number of PDF files on their site, including copies of every complaint filed against a registered owner or stable. Figure 13 shows the sites constituted of the highest percentage of PDF files.

2009		2010	
CA Agency	**PDFs**	**CA Agency**	**PDFs**
Regional Water Quality Control Boards	34849	Regional Water Quality Control Boards	29758
State Water Resources Control Board	34431	State Water Resources Control Board	29106
California Air Resources Board	30580	California Air Resources Board	28803
Department of Industrial Relations	27522	Department of Industrial Relations	26580
Board of Equalization	18577	Fair Political Practices Commission	18967
California Courts	14187	Board of Equalization	16560
Department of Transportation (Caltrans)	11388	California Courts	14174
California Energy Commission	9661	Department of Fish and Game	11928
Franchise Tax Board	9273	Department of Transportation (Caltrans)	11540
Department of Fish and Game	9215	California State Science Fair	10210
California State Controller	8317	Franchise Tax Board	9286
Department of Pesticide Regulation	7763	Horse Racing Board	8798
Horse Racing Board	7755	Division of Labor Statistics & Research	8536
Division of Labor Statistics & Research	7735	Department of Pesticide Regulation	8284
California State Science Fair	7440	Central Valley Bay-Delta Branch	8078
Department of Toxic Substances Control	7103	Department of Toxic Substances Control	7567
CA Dept of Corrections and Rehabilitation	6032	Secretary of State	7102
Secretary of State	5934	Managed Risk Medical Insurance Board	6812
CalPERS	5839	California Energy Commission	6572
Fair Political Practices Commission	5632	CA Dept of Corrections and Rehabilitation	6387
State Treasurer	5429	State Treasurer	6269
Managed Risk Medical Insurance Board	5358	California State Parks	5916
Department of Social Services	4759	State Park and Recreation Commission	5894
California Elections and Voter Information	4446	California Office of Historic Preservation	5880
Office of the State Fire Marshal	4178	CalPERS	5827
Department of Water Resources	3821	California Elections and Voter Information	5450
Department of Mental Health	3672	Museum Resource Center	5229
CA Community Colleges System Office	3490	Department of Social Services	5166
CA Integrated Waste Management Board	3411	Department of Water Resources	4839
Department of Corporations	3213	CA Integrated Waste Management Board	4716

FIGURE 12: # OF PDF FILES CAPTURED, 2009 & 2010

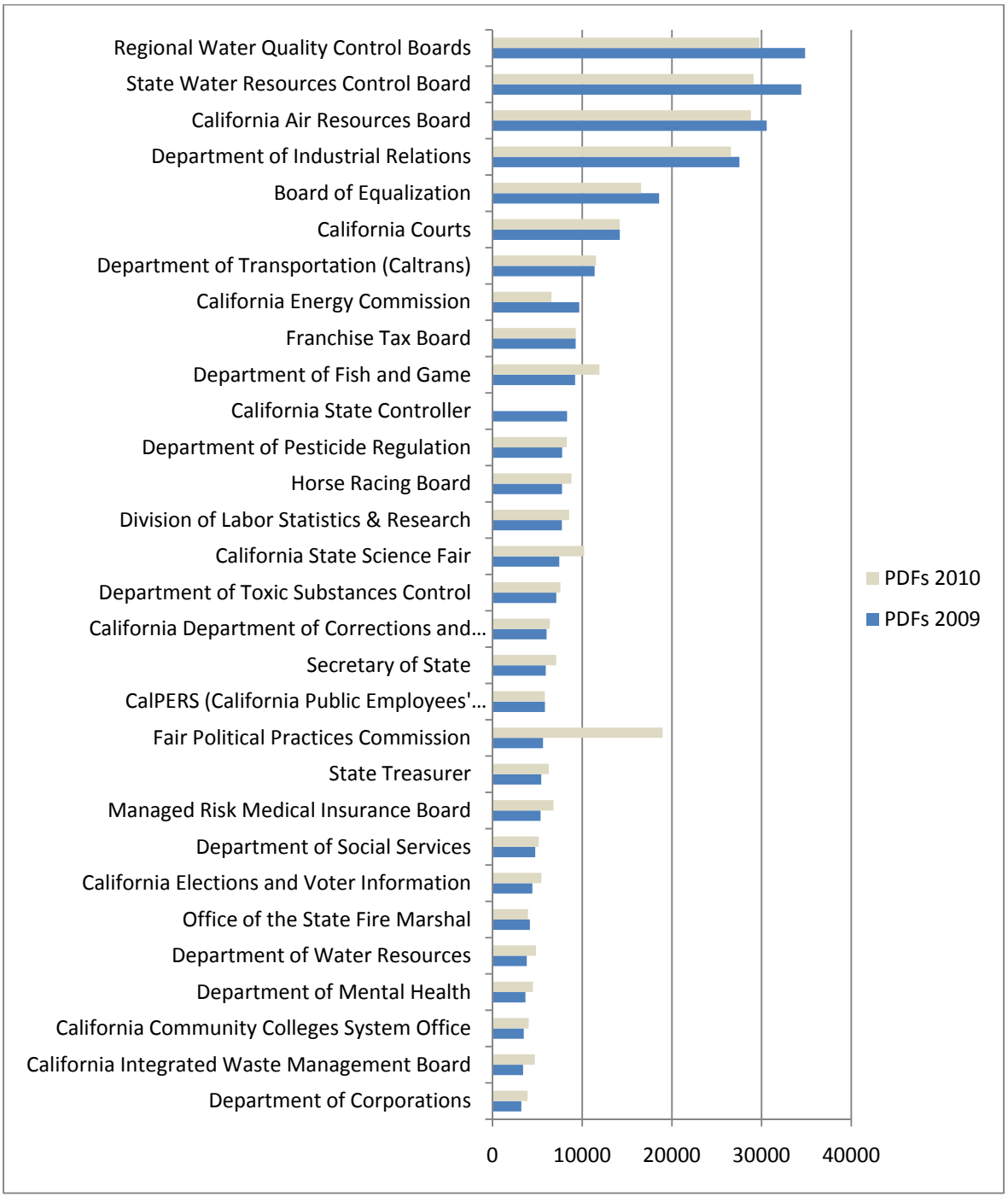

FIGURE 13: CA AGENCIES BY NUMBER OF PDF FILES, 2009

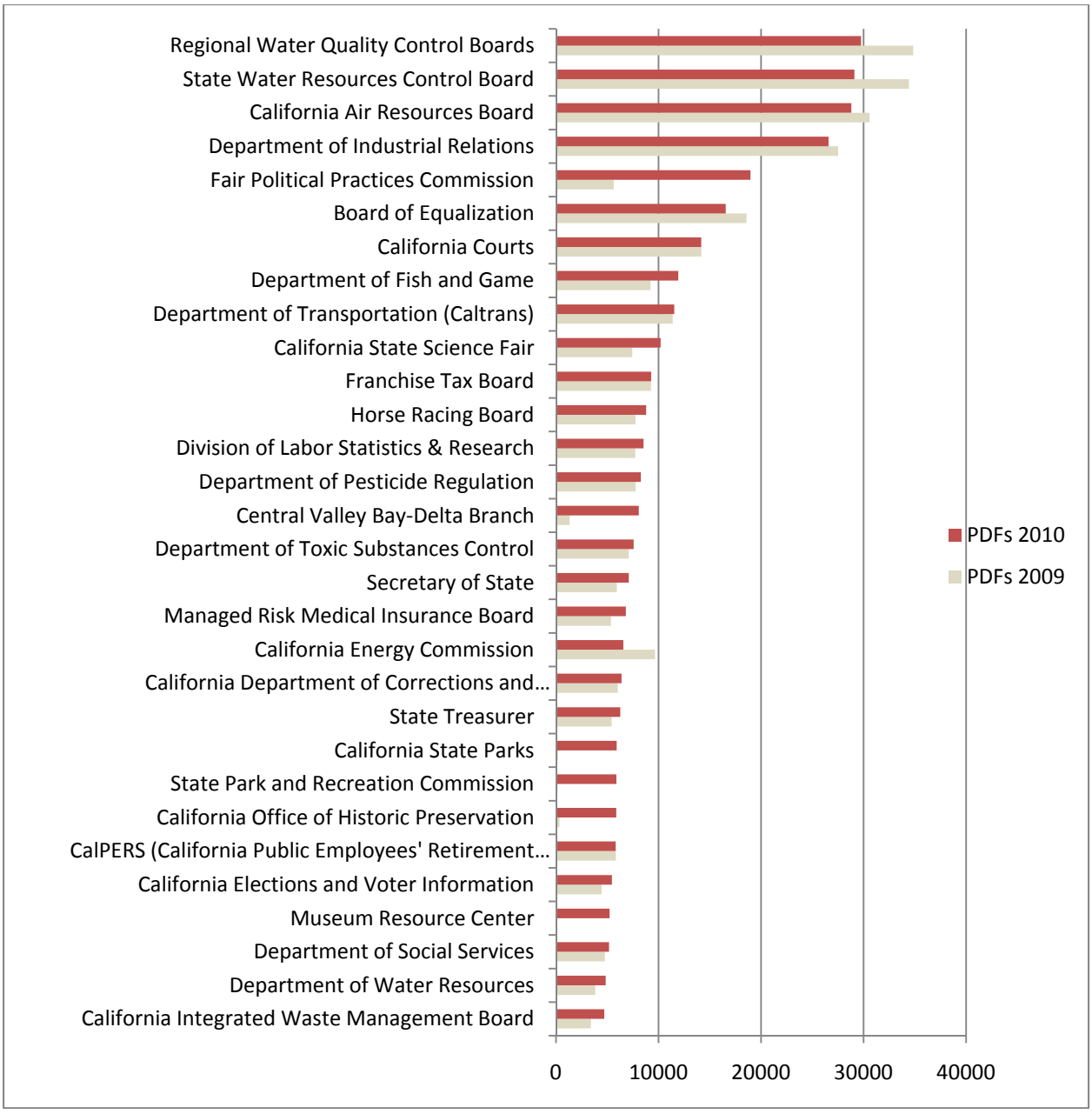

FIGURE 14: CA AGENCIES BY NUMBER OF PDF FILES, 2010

agency	pdfs2009	pdfs2010	% PDF loss
California State Controller	8317	0	100
Coastal Conservancy	981	0	100
Operations Control Office	910	0	100
Office of Lieutenant Governor	597	0	100
California Workforce Investment Board	285	0	100
California Department of Education	271	0	100
Office of Information Security and Privacy Protection	181	0	100
Cal Atlas: Geospacial Clearinghouse	62	0	100
Electricity Oversight Board	31	0	100
State Bar of California	1	0	100
Attorney General	1	0	100
CalGOLD (Business Permit Information)	1	0	100
Buy California Bonds	1	0	100
California Women's Commission	193	1	99.5
California State Railroad Museum	9	1	88.9
Santa Monica Mountains Conservancy	2287	404	82.3
Office of the State Public Defender	8	2	75
California Natural Resources Agency	1335	531	60.2
Departmento de Vehiculos Motorizados	197	81	58.9
Tax Service Center	2	1	50
California Business Portal	2	1	50
Bureau of Security and Investigative Services	171	108	36.8
California Emergency Management Agency	2218	1413	36.3
California Energy Commission	9661	6572	32
Office of the Patient Advocate	263	179	31.9
Board for Geologists and Geophysicists	283	203	28.3
Sierra Nevada Conservancy	447	324	27.5
Wildlife Conservation Board	23	17	26.1
Postsecondary Education Commission	1276	1002	21.5

FIGURE 15: HIGHEST PERCENTAGE OF PDF FILES LOST 2009-2010

agency	docs2010	pdfs2010	% pdfs
Office of the State Fire Marshal	4219	3953	94
Horse Racing Board	9610	8798	92
Department of Social Services	5678	5166	91
CA Alcoholic Beverage Control Appeals Board	1934	1752	91
Physical Therapy Board of California	1341	1188	89
Managed Risk Medical Insurance Board	7742	6812	88
Governor's Office of Planning and Research	560	476	85
Children and Family Services Division	2100	1752	83
Gambling Control Commission	1523	1253	82
Board of Equalization	20529	16560	81
High-Speed Rail Authority	2708	2187	81
CSU Board of Trustees	1191	956	80
Office of Administrative Law	1132	910	80
State Water Resources Control Board	36688	29106	79
California Law Revision Commission	3752	2958	79
Fish and Game Commission	1681	1319	78
Respiratory Care Board	1078	843	78
California Elections and Voter Information	7079	5450	77
Board of Pharmacy	2338	1805	77
Central Valley Flood Protection Board	2350	1813	77
Delta Vision	1392	1072	77
Commission on Judicial Performance	276	212	77
California Education Audit Appeals Panel	120	91	76
Reporting Transparency in Government Website	4089	2998	73
Fair Political Practices Commission	26269	18967	72
California Agricultural Labor Relations Board	2092	1498	72
Santa Monica Mountains Conservancy	571	404	71
Regional Water Quality Control Boards	42709	29758	70
Department of Child Support Services	3727	2592	70
Board of Forestry and Fire Protection	458	320	70

FIGURE 16: AGENCY SITES COMPOSED MOSTLY OF PDF FILES

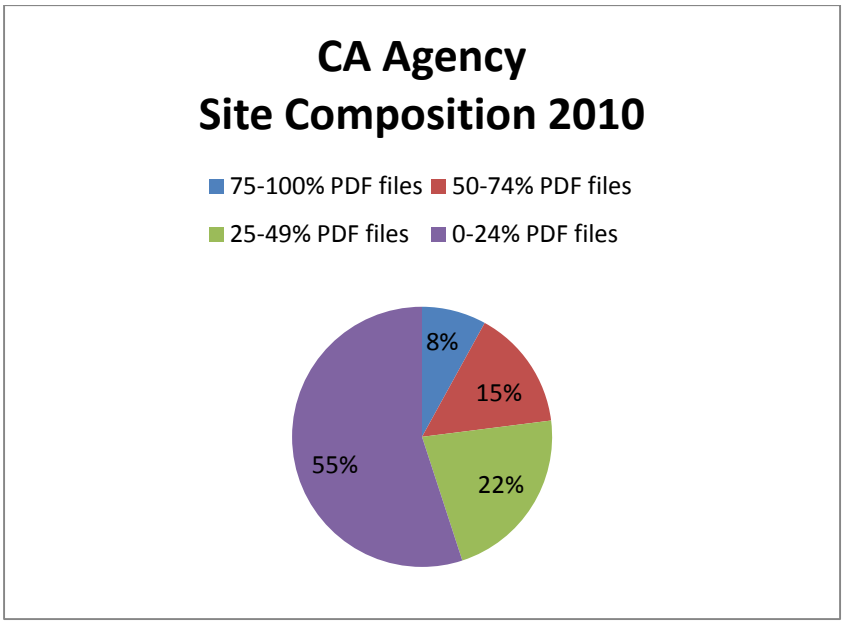

FIGURE 17: PERCENTAGE OF PDFS ON CA AGENCY SITES, 2010

The entire State of California Government web presence consists of 20.9% PDF files in Jan 2010.

COMPARISON BY STORAGE SIZE 2009 & 2010

The following table compares the top 30 sites captured by the storage size used, rather than by the number of documents. For 2009, 13 of the largest sites are not necessarily the sites with the greatest number of files. For 2010, 10 of the largest sites are not necessarily the sites with the greatest number of files. If sites in this table don't include a large overall number of files, they may include larger number of complex multimedia files.

2009		2010	
CA Agency	**bytes**	**CA Agency**	**bytes**
Regional Water Quality Control Boards	41323056369	Governor's Office	42094936958
State Water Resources Control Board	39854973329	State Water Resources Control Board	39724887397
California Air Resources Board	25701467809	Regional Water Quality Control Boards	29315074491
Department of Water Resources	16182044915	California Air Resources Board	26199300506
Department of Transportation (Caltrans)	14201801284	**Department of Consumer Affairs**	18376563303
Department of Industrial Relations	10827925798	Department of Transportation (Caltrans)	17513285467
California Energy Commission	10727432368	**Commission on Teacher Credentialing**	14516612738
Commission on Teacher Credentialing	10623094055	CA Integrated Waste Management Board	13705598969
CA Integrated Waste Management Board	10199635001	Department of Water Resources	13298383810
Department of Fish and Game	8183637230	Department of Industrial Relations	9956102653
Department of Pesticide Regulation	6384941421	California Energy Commission	9555258228
Floodplain Management	6267760681	Department of Fish and Game	9389973981
Department of Consumer Affairs	5294477346	California Department of Food and Agriculture	8097041299
Central Valley Bay-Delta Branch	4597049192	Central Valley Bay-Delta Branch	8067654221
California Courts	4488847283	**Central Valley Flood Protection Board**	7664100970
Department of Toxic Substances Control	4335293740	California State Parks	7116839443
California Emergency Management Agency	4327174007	State Park and Recreation Commission	7051314859
Board of Pharmacy	4161871126	**Department of Pesticide Regulation**	6871947674
Board of Equalization	3898526860	California Office of Historic Preservation	6862435969
Fish and Game Commission	3892288706	Museum Resource Center	6473639207
California Law Revision Commission	3625330903	**Floodplain Management**	6287674523
Office of Technology Services	3620517501	**CA Environmental Protection Agency (Cal/EPA)**	6116033423
Secretary of State	3436464773	**High-Speed Rail Authority**	5978641741
California State Lottery	3388334904	Assembly Democratic Caucus	5396201177
CA Dept of Food and Agriculture	3244574423	California Courts	4995002377
California State Assembly	3167713125	**Secretary of State**	4818673762
Office of Statewide Health Planning and Dev	2970678246	**Reporting Transparency in Govt Website**	4482306137
Assembly Democratic Caucus	2854064647	Board of Equalization	4149011056
Department of Mental Health	2589543115	Department of Toxic Substances Control	3957397572
Labor Market Information Division	2469595069	**Fish and Game Commission**	3954359232

FIGURE 18: CA STATE AGENCY SITES BY STORAGE SIZE: 2009 & 2010

CONCLUSIONS

ARCHIVING THE CALIFORNIA STATE GOVERNMENT WEB IS AN ACHIEVABLE GOAL

The Web Archiving Service has demonstrated that the State of California sites can be effectively captured, even though it is such a formidably large Web domain. The practice of building this archive has helped the CDL Web archiving team optimize the service for large-scale production use. The Service can capture not only of hundreds of sites simultaneously, but also sites that are rich with data sets, publications and other media content.

The largest barriers to capturing the State Government Web presence are issues of policy, not technical issues.

WEB ARCHIVISTS AND STATE AGENCIES NEED TO BEGIN STRATEGIC COMMUNICATION AND COLLABORATION

The State of California Web presence is obviously guided by a set of clear guidelines that is centrally administered. There is strong consistency in the appearance, navigation and imagery of State agency websites, and there is a growing consistency to the server restrictions imposed on those sites. These settings form the greatest barrier to the ability to archive and provide lasting access to this content. We need to reach out and begin strategic communication with the organization shaping the State's Web publishing practices to encourage more open access to Web content and, at the very least, to expressly permit this project's crawlers to archive State Web content.

As mentioned earlier, the Section 108 Study Group's recommendations include the assertion that Libraries and Archives should not be prevented from capturing the Web content of government agencies or political candidates. This content both reflects and shapes public policy in the State of California. It documents our political and cultural heritage for future researchers, and its creation was funded by California taxpayers. A policy of open access for Web capture would be well in keeping with the State's policies on transparency in government. Barring that, there are steps that content owners can take to make their sites preservation-ready for CDL while still addressing the bandwidth concerns their system administrators might have. The following two lines in a robots.txt file will make it possible for CDL to archive the site, but will apply existing restrictions to other crawlers:

```
User-Agent: cdlwas_bot
Disallow:
```

THE ARCHIVE PROVIDES VALUE TO THE STATE OF CALIFORNIA AGENCIES

In an environment of fiscal challenges, most State agencies do not have the resources of maintain an archival record of their digital publishing activity. As the composition of State government changes; as agencies disappear or are merged with other agencies, the California State Government Web Archive is able to provide a lasting record of State publications both to the general public and to the agencies themselves.

If collaborative efforts were to go beyond simply allowing permission to capture, the archive could become even more valuable to the State agencies. The National Archives has established just such a collaborative relationship with the U.K. Government Agencies, and serves as the archive of record for U.K. government Web content. In addition to configuring servers to permit archive crawlers to capture content, the agencies also configure their servers to point back to the archive when users encounter 404 errors on their sites. The government agency sites are linked to the archive, so that users do not need to know about the archive in order to benefit from it, and the agencies are able to provide seamless service to the public for both current and archival information.[7] This kind of collaboration would be a tremendous benefit to California citizens and researchers.

UC Librarians may want to revise the distribution of the priority agency sites

At first glance, it appears that the watch-list of agency sites is not evenly distributed based on the size of each website. Figure 2: "Distribution of Work by Campus" shows that UC San Diego and UCLA are monitoring the largest sites in the list. This may or may not be an issue, depending on the actual resources available at each library and the relative importance of the sites in the list. The librarians in this group may want to examine a couple of the charts closely to see if they surface any websites that deserve more frequent capture. Figure 4: "Comparison of 2009 and 2010 Top 30 CA Agencies by Number of Files", and Figure 9: "Number of PDF Files Captured" may be of particular use.

There are still unexplored possibilities for Web archive analysis and QA tools

Most of the statistics and graphs in this report were produced using custom queries against the archive's database. These reports are not currently available in the WAS curatorial interface, but are relatively easy to produce on-demand. Once the content is in an archive, and reports on the archived contents can be queried in a database, one can see things about that Web content that cannot be discerned on the live Web. It is easy to tell which sites really are the largest or the most volatile and which offer the most multimedia content. These statistics also suggest that tools can be developed to alert a curator to anomalies where there are hundreds of sites being monitored. There are also certainly more questions that can be asked of this content that have not been asked in this study.

[7] Amanda Spencer and Alison Heatherington. "Continuity and Preservation: The National Archives approach to maintaining permanent access to the web presence of UK Central Government" http://www.netpreserve.org/events/active_solutions/11_IPRES-IIPC09%20TNA%20presentation3.pdf