

IT TAKES A VILLAGE TO SAVE THE WEB: THE END OF TERM WEB ARCHIVE

Tracy Seneca, Abbie Grotke, Cathy Nelson Hartman, Kris Carpenter

The End of Term Web Archive collaboration began in May of 2008, when the Library of Congress, the Internet Archive, The University of North Texas, the California Digital Library, and the U.S. Government Printing Office agreed to join forces to collaboratively archive the U.S. government web. The goal of the project team was to execute a comprehensive harvest of the federal government domains (.gov, .mil, .org, etc.) in the final months of the Bush administration, and to document changes in the federal government websites as agencies transitioned to the Obama administration.

This collaborative effort was prompted by the announcement that the National Archives and Records Administration (NARA), which had conducted harvests of prior administration transitions, would not be archiving agency websites during the 2008 transition.¹ This announcement prompted some considerable debate about the role of NARA in web archiving and the value of archiving websites in their totality. It also came just as the International Internet Preservation Consortium (IIPC) held its 2008 General Assembly. All five project partners are IIPC members, and were able to convene an immediate meeting to discuss what actions should be taken. With little time and no funding, the five End of Term (EOT) Project organizations responded together with the range of skills and resources needed to build the archive.

The End of Term Web Archive (eotarchive.cdlib.org) includes federal government websites in the legislative, executive, and judicial branches of government. It holds over 160 million documents harvested from 3,300 websites, and represents sixteen terabytes of data. This article

will outline the steps taken to build the archive, detail the innovations that made the project successful, and will convey plans for the forthcoming 2012 End of Term collection.

BACKGROUND

As stated above, all five EOT partners were already active in the IIPC, an organization with a strong role in helping national libraries harvest and preserve their nation's web publications.² Most partners either had the direct capacity to archive web content, or had support for doing so.

The Internet Archive (IA) has been harvesting and providing access to web content since 1996. They conduct their own broad crawls of the internet (www.archive.org), they host the Archive-It service to enable other organizations to archive materials, and they also provide services for large-scale, comprehensive captures of specific web domains.³ The IA had previously conducted the 2004 End of Term Harvest in partnership with NARA, and has also collaborated with other libraries, archives, and memory institutions to preserve web content of national importance.

The Library of Congress has been curating thematic collections of web content since 2000, including United States National Elections, the Iraq War and the Events of September 11, 2001.⁴ The Library also enabled organizations nationwide to begin archiving web-based materials through its National Digital Information and Infrastructure Preservation Program (NDIIPP), which funded a number of web archiving initiatives including the Web-at-Risk, The Web Archivists' Workbench and the K-12 Web Archiving program.⁵ The Library continues to play a key role in promoting internet preservation through its National Digital Stewardship Alliance program.⁶

The University of North Texas (UNT) is also among the early leaders of web archiving, having founded the CyberCemetery in 1997. The CyberCemetery provides permanent public access to the web sites and publications of defunct U.S. government agencies and commissions.⁷ In addition to having the capacity to conduct large-scale web harvests locally, UNT has contributed extensively to the assessment and study of web archive curation activities, and has actively contributed to tools in support of web archiving.

The California Digital Library (CDL) is engaged in large-scale web harvesting, both in support of collections for the University of California and on behalf of other organizations. CDL hosts the Web Archiving Service, which enables librarians and archivists to archive websites, and was funded in part by the Library of Congress NDIIPP program.⁸ CDL has been actively engaged in shaping web archiving standards and in helping to integrate web content selection into bibliographer workflows.

While the Government Printing Office (GPO) was not directly engaged in web crawling when the End-of-Term project began in 2008, GPO administers the Federal Depository Library Program and has a strong interest in government publications. They joined project calls to stay informed about the work.

Among them, these five organizations have been directly involved in content selection and curation, in shaping policies around web archiving, in shaping the underlying standards and tools that make web archiving possible, and in designing and building technology to support web archiving. All of these strengths were drawn upon in the course of the project.

SELECTION AND CURATION

The 2008 End of Term harvest did not start with a blank slate; a number of web archiving efforts had previously captured U.S. federal government domains, and there were additional sources of information for identifying the websites within the scope of the federal government domains.

The project began by compiling the known sources for a seed list to build the archive.

A seed list is a list of URLs that very often corresponds to the home page address for particular websites. These are the URLs that a web crawler takes as the starting point to direct the harvest of any given group of websites. Some websites are complex enough to require more than one starting point, and some government agencies are complex enough to have a series of seed URLs associated with them. The National Institutes of Health are a good example of that; within the “nih.gov” domain, there are additional subdomains, such as “nhlbi.nih.gov”, along with related domains such as “cancer.gov”. The project partners had a few existing sources to draw from, including a list derived from the usa.gov website and a seed list used in Stanford University’s WebBase project.⁹ Note that all of these lists showed some variation in the number of “websites” considered to constitute the U.S. government web presence. The 2011 *State of the Federal Web Report* identifies 1,489 domains and 11,013 distinct public websites in the executive branch alone.¹⁰ The 2011 report represents the first detailed survey of web domains that some federal agencies have done, so it certainly unearthed sites that were not identified in any of the 2008 lists. However, despite their familiarity, the definition of a “website” can be hazy; many of those 11,013 websites would have been included in the 2008 End of Term web archive as components of larger sites.

The lists assembled only offered a minimum degree of metadata. In many cases the site name associated with a URL was available, but in some cases no additional metadata at all was available. Furthermore, these seed lists had been used in other projects with a sometimes significantly different scope. There were occasional “.gov” addresses pertaining to U.S. states, and there were also older URLs that were no longer functional. The lists also included a great many sites that are owned by the federal government, but that employ domains beyond “.gov”, including “.mil”, “.org” and even “.com”. UNT served as the gathering point for these lists, assembling them into a database built for the project.

Owing to the quality issues with this collection of seed lists, the project team agreed that some degree of curation was needed before the list could be used to run the first harvest of sites. The project team also agreed that the job was more than any one organization could take on, and that this could be a good opportunity to “crowdsource” the project to a wider community of government information specialists. Working with input from the Library of Congress and CDL, UNT built the URL Nomination Tool, which would allow volunteers to review the list of URLs and mark them as either in-scope or out-of-scope for the project. By marking items out-of-scope, a curator could indicate that the URL was either no longer functional or that it was not a federal government site. By marking an item in-scope, a curator could indicate that the site would be particularly vulnerable to significant change during the transition to a new administration, and that it should be harvested with greater frequency and depth. Seed URLs were given one in-scope point for each source list they appeared in, so for example, if www.whitehouse.gov appeared in all five source lists, it got an in-scope score of “5,” which could be increased or decreased by the curators’ votes. Curators could also add URLs that did not appear in the combined list.

The Nomination Tool also enabled curators to supply metadata for seed URLs including site name, agency name, branch of government, and comments that might be useful for the crawling team. Curators could search for specific URLs or browse URLs alphabetically by domain and subdomain. (See Figure 1)

The Nomination Tool was introduced at the 2008 ALA Annual Conference GODORT meeting, and further outreach was conducted to draw participation from email lists. Participation, however, was relatively low; the Nomination Tool ultimately held 4,622 URLs, with participation from twenty-nine nominators. Curators voted on approximately 500 of the URLs in the tool. It is possible that by providing so many URLs by default, curators interpreted the list as being complete, and were less inclined to provide votes or metadata. Even so, the Nomination Tool became a critical source for all partners to draw upon, both in the harvesting phase, and in the construction of the archive access gateway.

THE HARVEST

As the harvest of web content began in early September 2008, each contributing partner envisioned slightly different roles for collection building. The IA would conduct broad crawls of all of the seeds in the list, “bookmarking” the collection with comprehensive crawls at the beginning and end of the project. It was not possible for the IA to crawl these sites continuously over the course of the year-long collection phase, so IA, CDL and UNT ran shorter duration crawls between the book-end crawls at key intervals, i.e. pre-election, post-election, pre-inauguration, and immediately post-inauguration. The Library of Congress, meanwhile, focused on legislative branch websites, so collection was staggered across institutions both along timelines and in the focus of content collected. It was further envisioned that CDL would focus

on sites based on selection activity in the Nomination Tool, but ultimately they decided to also use the entire seed list for its harvest. UNT focused their collecting on agency sites defined by their government information specialists as meeting the requirements of their collection development policies. Crawl timelines are charted in Figure 2.

The copyright and intellectual property issues surrounding the emerging field of web archiving also came into play in this project. All project partners who crawled websites chose to ignore “robots.txt” files in their harvest settings. A robots.txt file is a set of instructions that web server administrators can provide to direct crawler activity on their websites. It is commonly used to prevent a crawler, such as the Google index crawler, from using unnecessary bandwidth on gathering image or style sheet files not needed to effectively index a website. It can also be used to explicitly direct crawlers not to collect any content whatsoever.

The use of robots.txt directives on public domain government websites to prohibit the archiving of taxpayer-funded content is controversial. The project partners considered all sites within this harvest to be within the public domain, and so ignored robots.txt instructions prohibiting capture. This decision is further supported by the Section 108 Study Group, convened in 2006 to consider revisions to copyright exceptions for libraries in keeping with advances in digital media. The Section 108 Study Group unanimously recommended that Federal, State and local government entities should not have the right to opt out of having their publicly available content archived by libraries.¹¹

THE DATA

After crawling activity ended in the early fall of 2009, portions of the data were distributed among the EOT partners; the Internet Archive held 9.1 TB of data, CDL held 5.7 TB, and the

UNT held 1TB. The next task was to assemble all of the archived content in one place, and to ensure that at least one full copy of the entire archive was held in a geographically separate location. This phase of the project was led by the Library of Congress, and required that all partners make their EOT content available for transfer. The Library of Congress then made the aggregate data set available to the partner libraries. The challenges of this task are described in detail in *The End-of-Term Was Only the Beginning*, a Signal Blog article on the project.¹² The EOT partners employed “Bag-It”, a data transfer standard developed under the NDIIPP program to support the transfer of grant-funded content to the Library of Congress.

The NDIIPP grant projects gave rise to data transfer innovation with the Bag-It specification, and the scale of data transfer in the End of Term project prompted use of such technical innovations to support large scale data transfer. NDIIPP had developed tools to support the Bag-It specification: Bagit Library, a Java based, Unix command-line tool for making, manipulating, and transferring, and validating bags over the network, and a Bagger desktop tool for working with bags.¹³ The Library used the Bagit tools for transferring content, and the tools were available for partners to bag their content and make it available for transfer. Starting in May 2009 and running for the course of about a year, all content was transferred to the Library of Congress. The Library then provided CDL and UNT content to the Internet Archive, and UNT received a full copy of the entire data set.

The End of Term Archive provided important data for researchers. In 2009, UNT received Institute of Museum and Library Services (IMLS) funding for a research project called *Classification of the End-of-Term Archive: Extending Collection Development to Web Archives*.¹⁴ Recognizing that librarians would need the capability to identify and select materials

from Web archives in accord with collection development policies and to then characterize these materials using common metrics that demonstrate their value, the project investigated innovative solutions to address these needs in two work areas.

Work Area 1 addressed archive classification. Classification of the EOT Archive involved both structural analysis and human analysis. Link analysis, cluster analysis, and visualization techniques identified the organizational and relational structure of the EOT Archive and produced clusters of related websites from a representative set of the Archive's URLs. The project's subject matter experts (SMEs) classified the same set of URLs according to the SuDocs Classification Scheme using a Web-based application developed by project staff. The resulting classification served as the standard against which the effectiveness of the structural analysis was evaluated. As an additional exercise to test the topical relatedness of the clusters' members (i.e., Websites), a tool was developed to allow the project's SMEs to add subject tags to each cluster. Comparisons showed that the automated clustering processes were significantly successful in grouping topical areas as shown by comparison to the SME tagging.

Work Area 2 focused on web archive metrics. Identification of metrics for Web archives was informed by the project's SMEs who participated in two focus groups to identify and refine the criteria libraries use for acquisition decisions. UNT conducted a review of existing statistics and measurements used by academic libraries and identified content categories for the EOT Archive. This work culminated in a proposed set of web archiving metrics which was then submitted to an International Standards Organization (ISO) working group currently analyzing the same issues. This ISO working group (ISO TC46 SC8 WG9) is currently preparing a technical report, and the UNT research team met twice with the working group's chair to review the proposed metrics.

Anticipating researchers' needs to understand the scope and type of content in the Archive, UNT analysts also investigated which data elements could be readily extracted from the Archive's files. Further research in this area will continue in the coming year.

In addition to helping clarify the scope and value of the materials in the End of Term Web Archive, this work also highlights the complex and rich role of web archives in library collections. Beyond providing passive "replay" of web content as it appeared in the past, web archives may also serve as dynamic sources for data analysis, and can enable discoveries that were not possible when that same content was only available on the live web.

THE ARCHIVE

The End of Term project partners intended from the start of the project to make the resulting archive freely available to the public. Once the data transfer work was complete in mid-2010, the work of providing public access could begin. IA and CDL agreed to collaborate on a public access portal to the copy of the data held at the IA. While both organizations provide public access systems for web archives, the End of Term content still posed a challenge. The captures of web content were run outside of the context of Archive-It and the Web Archiving Service, which meant that the content couldn't be delivered via either of the well-established discovery systems that both services offer. The team also wanted to be able to provide more than just URL lookup or full text search functions. There were three distinct challenges to providing public access: the development of a portal interface for browsing a site list, the significant task of indexing nearly sixteen terabytes of data, and the delivery of rich data visualization tools enabling researchers to better understand the scope of the archive.

The portal (eotarchive.cdlib.org) uses CDL's eXtensible Text Framework (XTF) to provide faceted browsing and metadata search of the site list, drawing on metadata records extracted by IA. XTF is an open source digital library platform that is commonly used to provide access to digitized images and documents, and is the technology behind CDL's Online Archive of California and eScholarship Repository and has been used by a range of organizations beyond CDL.¹⁵ The default open source version supports processing of PDF, EAD, NLM, Dublin Core and TEI formats, and includes a book reader; it would not appear at first to be an obvious tool for web archive discovery. However, XTF can be easily configured to process other metadata formats, and the content itself does not have to be co-located with the metadata files.

The Internet Archive had previously built a MODS-extraction tool to generate basic metadata records from a seed list, and using that tool, they ran a record set based on the End of Term sites. Initially, CDL tried these MODS records in XTF on an experimental basis, and the results worked well enough to show that XTF would be a viable option. The MODS format was less ideal than simple Dublin Core, so CDL and IA worked together to produce Dublin Core records for the End of Term sites. The metadata elements are described in Figure 3.

The coverage and source elements provide the data needed to explore the archive by government branch or URL segment. The title and provenance elements allow the user to search the site list by site name or URL. While HTML title tags can be notoriously unreliable, the government sites did tend to have useful titles; only about 200 sites out of over 3,300 lacked title information. The abstract provides information on the brief display of the site records, and the identifier links the user through to the displayed page at the Internet Archive. The automatic extraction of subject

terms from the seed list unfortunately did not work well for discovery; many subject terms were used only once so that they did not tend to lead the user to related materials.

The success of this approach has promising implications. The IIPC has long sought a means to collaboratively build archives on topics of international importance. Experiments are currently underway to build a distributed collection of 2012 Olympics web archives, and the End of Term archive demonstrates that the discovery interface and content can be at separate and even multiple locations. This approach also holds promise for integrating web archived content with topically related scanned materials – e-books, documents and imagery. Very few archives currently do this; the UCLA Campaign Literature Archive is a rare exception and an important one.¹⁶ Web archived materials are stored in a unique format that requires additional software to ‘replay’ the archived site. This poses a challenge for archive display, and can lead to unnecessary silos of information. It should not matter to an end-user how materials in an archive were acquired. Regardless of whether it was scanned or harvested, the content itself is what matters. The potential exists to use XTF, Omeka or other discovery platforms to aggregate access to multiple web archives or to integrate web archived content with more traditional digital formats.

The full text search of the End of Term Web Archive presented an entirely different problem.

Web archiving technology has been at an important crossroads in 2010 and 2011, as organizations engaged in large scale archiving have determined to migrate to more powerful indexing tools. Thus far, most web archives have relied on Nutch, an open-source, Lucene-based full text search engine. Nutch has fallen short in many respects, and the open source community is instead devoting more attention and development to SOLR. SOLR is a widely adopted full text search engine, also built on Lucene, and is used by hundreds of libraries and archives around the

globe to search metadata as well as the full text of digitized books and other resources.

Programmers have adapted SOLR for web archives on an experimental basis at a number of libraries including the British Library and at CDL, and an increasing number of web archives will transition to SOLR for public access searching over the course of the next year. In this interim phase, the full text search service deployed for the EOT 2008 archive was generated using TNH, a custom packaging of Lucene with extensions for support of web archives.¹⁷

When searching the full text of the End of Term Web Archive, the first round of results will list the most relevant result from each website in the result set. You can select “More from [this site]” to view the remaining results from any given site. Because TNH is being used while the open source community migrates to SOLR, full text search features for the End of Term Web Archive are likely to improve when SOLR is robust enough to support the demands of large scale web archives.

The capacity to generate enhanced discovery tools and data visualizations for web archives is also at a turning point, based on newly emerging standards and tools at the Internet Archive. The project team is working on exposing a series of visualizations enabled by Google Analytics in the browser, via CoolIris for navigating the collection by image, and via open source link graph and analysis tools. The key to enabling these alternative views of the archive is the introduction of the Web Archive Transformation (WAT) specification for structuring metadata generated by web crawls.¹⁸

Web crawlers do not return with mirrored copies of websites, but instead return with large container files called WARCs, which hold both the content of thousands of files and metadata about those files.¹⁹ While WARC files enable web archives to more easily manage the massive

scale of storage required, they also pose challenges for indexing and analysis tools. WAT utilities extract the metadata stored in WARC files into a highly optimized form that can be analyzed in a distributed processing environment such as Hadoop.²⁰ WAT has been quickly adopted for experimental work at many organizations involved in web archiving, including project partners UNT and CDL. More of these WAT-enabled visualization services will be released in 2012 as we build toward the development of the next End of Term Archive.

THE NEXT ARCHIVE

The End of Term project has resumed for an End of Term 2012-2013 archive, and help is needed to identify websites for collection, particularly those that might be most at-risk of change or deletion at the end of the presidential term. Nominations of any U.S. federal government domains are welcome. Based on what was learned from the 2008-2009 archive project, the project team has also identified a few topical areas needing focused effort by subject experts, including but not limited to:

- *Judicial branch websites.

- *Important content or subdomains on very large websites (such as NASA.gov) that might be related to current presidential policies.

- *Government content on non-government domains (.com, .edu, etc.).

Volunteer nominators will be asked to contribute as much time and effort as they are able, whether it be a nomination of one website or 500 websites. Nominators will be given access to the Nomination Tool, updated for the 2012-2013 project.

Government document experts, subject experts, and any others interested in helping identify U.S. federal government websites for collection and preservation are encouraged to contact the project team at eotproject@loc.gov.

The project team plans to focus on recruitment of volunteer nominators in the summer of 2012. In July or August 2012, a baseline crawl of government web domains will begin. The focused crawling by partners will occur mostly in the fall of 2012, with partners crawling various aspects of government domains at varying frequencies, depending on selection policies and interests. At that time, the team will also determine a strategy for crawling prioritized websites. The crawls will continue into 2013, with a final crawl date depending on the outcome of the election.

SUMMARY AND CONCERNS

The ad-hoc collaboration that came together in response to the impending transition of presidential administrations in 2008 has been highly successful. The existing EOT partners have moved forward on the 2012 archive without hesitation, and Harvard University Library has joined the EOT partnership. The project has made use of emerging tools, and has in some cases driven the development of tools and practices that have since been more widely adopted. The 2008-2009 federal government content itself is now held at three institutions, all of which have robust digital preservation practices in place. That content has already supported grant-funded research activity, and will likely support further research and analysis in the future.

While successful, the EOT partners agree that there is still cause for concern, some of which is evidenced in the *2011 State of the Federal Web Report* mentioned earlier in this article. Without a comprehensive inventory, some websites were likely missed in the 2008 archive. A larger risk,

however, is the assumption that the change of administrations is the most meaningful indication of risk for widespread change or loss in web-based government publications, and that an archive collected every four years will be sufficient. The mid-term elections of 2010 are widely considered to be as consequential as the 2008 election. Whether that is evident in the scope and content of the federal government web presence is not yet known, but the most significant triggers that should prompt preservation and archiving activity may not be as obvious as a shift in administrations or political parties. The Report notes that agencies “have plans to eliminate or merge a total of 442 domains, mostly in FY3 and FY4 of calendar year 2011”.²¹ This represents about 30 percent of the existing Executive Branch domains, and is prompted less by political change, than by an understandable effort to streamline and improve agency website management.

The survey behind the *State of the Federal Web Report* was conducted with fifty-six federal agencies in the fall of 2011. The report is very much focused on issues of design consistency, governance and content management; the aim is clearly to reduce the Federal Government “web footprint” and to make web communications more efficient. There is no indication that questions about preservation and archiving were included in the survey, and preservation is not addressed in the report. While some details are provided concerning the domains to be eliminated, each of which may represent many websites, there is no mention of whether any of these materials will be archived. Both Archive-It and the Web Archiving Service have partnerships with individual federal agencies to preserve the public record of their web publications; many others pro-actively contact UNT’s Cyber-Cemetery project when sites are to be decommissioned. Ultimately, agencies have as varied an approach to web content preservation as they do to publication and management.

The End of Term partners agree that a comprehensive archive of the U.S. federal government web presence should ideally be undertaken on a yearly basis. While the EOT partners assembled the resources to carry out a harvest with each presidential election, a more consistent and ongoing effort would require additional funding. In keeping with the spirit behind the *Federal Web Report*, EOT partners would very much like to see preservation and archiving become an assumed part of any effort to more consistently and effectively manage web-based government publications.

REFERENCES

1. National Archives, *Memorandum to Federal Agency Contacts: End-of-Administration web snapshot*, NWM 13.2008, www.archives.gov/records-mgmt/memos/nwm13-2008.html.
2. International Internet Preservation Consortium, netpreserve.org/about/index.php.
3. Archive-It, www.Archive-It.org.
4. Library of Congress Web Archives, lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html.
5. National Digital Information and Infrastructure Preservation Program (NDIIPP), www.digitalpreservation.gov.
6. National Digital Stewardship Alliance, www.digitalpreservation.gov/nds/.
7. CyberCemetery, govinfo.library.unt.edu/.
8. Web Archiving Service, webarchives.cdlib.org/.
9. Stanford Web Base, diglib.stanford.edu:8091/~testbed/doc2/WebBase/.
10. The .gov Reform Task Force, *State of the Federal Web Report*, 4, www.usa.gov/webreform/state-of-the-web.pdf. For more information about this effort, see the .gov Reform Effort: Improving Federal Websites, www.usa.gov/WebReform.shtml.
11. Section 108 Study Group Recommendations, 80,

www.section108.gov/docs/Sec108StudyGroupReport.pdf.

12. The Signal Blog, the Library of Congress, *The 'End of Term' Was Only the Beginning*, July 26, 2011, blogs.loc.gov/digitalpreservation/2011/07/the-end-of-term-was-only-the-beginning/.

13. Bagger download page at Sourceforge: sourceforge.net/projects/loc-xferutils/files/loc-bagger/.

14. EOTCD Project Website, research.library.unt.edu/eotcd/.

15. XTF – eXtensible Text Framework, xtf.cdlib.org/

16. UCLA Campaign Literature Archive, digital.library.ucla.edu/campaign/.

17. Lucene, lucene.apache.org/.

18. Web Archive Transformation format, webarchive.jira.com/wiki/display/Iresearch/Home.

19. WARC Standard, www.iso.org/iso/catalogue_detail.htm?csnumber=44717.

20. Hadoop, hadoop.apache.org/.

21. *State of the Federal Web Report*, 12.

Tracy Seneca, Web Archiving Service Manager, California Digital Library,

Tracy.Seneca@ucop.edu

Abbie Grotke, Web Archiving Team Lead, Library of Congress, abgr@loc.gov

Cathy Nelson Hartman, Associate Dean of Libraries, University of North Texas,


cathy.hartman@unt.edu

Kris Carpenter, Director, Web Group, Internet Archive, kcarpenter@archive.org

End of Term Presidential Harvest 2008 Other Projects | Help

[Project Home](#) [About This Project](#)

☆☆☆☆☆☆☆☆☆☆

URL
<http://www.change.gov> 

Domain SURT
[http://\(gov,change,](http://(gov,change,)

Nominations
2

Nomination Score
2

Nominated By
Abbie Grotke - Library of Congress
Rebecca Blakeley - McNeese State University

Comment
Obama-Biden Transition team website, launched November 5, 2008

Branch
executive

Title
Obama-Biden Transition Project

Related URLs
<http://change.gov>

Hosted by The University of North Texas Libraries
UNT and State of Texas: [UNT](#) | [UNT Search](#) | [UNT News and Events](#) | [State of Texas](#) | [State-wide Search](#)
Policies: [UNT Web Accessibility Policy](#) | [AA/EOE/ADA](#) | [Privacy Statement](#) | [Disclaimer](#)

UNIVERSITY OF NORTH TEXAS

Figure 1

	2008				2009				
	Sep	Oct	Nov	Dec	Jan	Feb	Mar-Apr-May	Jun-Jul	Aug-Sep
IA	Broad					Broad			Broad
LC			Legislative	Legislative	Legislative	Legislative			
UNT		Selected	Selected			Selected			
CDL		Broad				Broad	Broad		
IA					Prioritized URLs		Prioritized URLs		Prioritized URLs

Figure 2

DC Element	Notes
Title	Derived from the HTML <title> tag of the site's home page. This is not always correct and not always present.
Identifier	The archival URL of the site's home page at the Internet Archive.
Provenance	The original URL of the site on the live web. Source for URL lookup.
Date [1]	The first date of capture.
Date [2]	The last date of capture.
Description	Derived from HTML <meta> tags on the site's home page, when present.
Coverage	The branch of government, derived from metadata included with the seed list and provided in the UNT Nomination Tool. Shown in site list facets.
Source	This repeatable field provides individual segments of the site URL, such as "nasa" or "senate". Shown in site list facets.

Figure 3