

Web Archives and Large-Scale Data: Preliminary Techniques for Facilitating Research

TCDL
May 24, 2012

Nicholas Woodward
Latin American Network Information Center
nwoodward@mail.utexas.edu

presidencia.gob.hn

. . . During the Coup . . .

Test Page for the Apache HTTP Server o... +

← → ⌂ A http://wayback.archive-it.org/176/20090919203226/http://www.presidencia.gob.hn/ ☆ ↻ Google 🔍

You are viewing an archived web page, collected at the request of University of Texas at Austin Libraries, Latin American Government Documents Archive using [Archive-It](#). This page was captured on 20:32:26 Sep 19, 2009, and is part of the [Latin American Government Documents Archive, LAGDA](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. hide

Red Hat Enterprise Linux Test Page

This page is used to test the proper operation of the Apache HTTP server after it has been installed. If you can read this page, it means that the Apache HTTP server installed at this site is working properly.

If you are a member of the general public:

The fact that you are seeing this page indicates that the website you just visited is either experiencing problems, or is undergoing routine maintenance.

If you would like to let the administrators of this website know that you've seen this page instead of the page you expected, you should send them e-mail. In general, mail sent to the name "webmaster" and directed to the website's domain should reach the appropriate person.

For example, if you experienced problems while visiting [www.example.com](#), you should send e-mail to "[webmaster@example.com](#)".

For information on Red Hat Enterprise Linux, please visit the [Red Hat, Inc. website](#). The documentation for Red Hat Enterprise Linux is [available on the Red Hat, Inc. website](#).

If you are the website administrator:

You may now add content to the directory `/var/www/html/`. Note that until you do so, people visiting your website will see this page, and not your content. To prevent this page from ever being used, follow the instructions in the file `/etc/httpd/conf.d/welcome.conf`.

You are free to use the image below on web sites powered by the Apache HTTP Server:

[\[Powered by Apache \]](#)

presidencia.gob.hn

. . . After the Coup

You are viewing an archived web page, collected at the request of University of Texas at Austin Libraries, Latin American Government Documents Archive using [Archive-It](#). This page was captured on 20:32:11 Dec 19, 2009, and is part of the [Latin American Government Documents Archive, LAGDA](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. hide

[INICIO](#) [CONTÁCTENOS](#) [WEB-MAIL](#) [IR](#)



Presidencia de la República

[CASA DE GOBIERNO](#) [GOBIERNO](#) [NOTICIAS](#) [PRESIDENTE](#) [REVISAR CORREO](#) [SALA DE PRENSA](#) [VIDEOS](#)



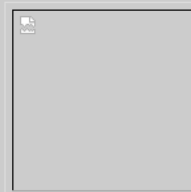
Micheletti: “Exigimos que nos respeten”

TEGUCIGALPA.- El Presidente Roberto Micheletti, exclamó que “exige” a la comunidad internacional que respeten las decisiones de Honduras como país soberano en un evento para conmemorar el 184 aniversario del Ejército Nacional.

“Quisiera pedirles a las naciones del mundo que respeten este pequeño país, aquí en Honduras, no tenemos dinero y no tenemos petróleo pero aquí [...]

[Ver más →](#)

“Mel” ya es historia



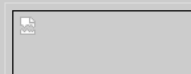
3, DICIEMBRE , 2009

TEGUCIGALPA.- El Presidente Roberto Micheletti dijo anoche que si el Congreso Nacional (CN) se reunió ayer para decidir sobre la restitución o no del depuesto mandatario, Manuel Zelaya, fue porque este así lo pidió, pues él siempre consideró que el tema debió pasar por una resolución de la Corte Suprema de Justicia (CSJ). Micheletti manifestó lo [...]

[Ver más →](#)

Cadena 19 de Noviembre

Parlamento alemán confirma apoyo a comicios



27, NOVIEMBRE , 2009

BERLÍN, ALEMANIA.

El parlamento alemán decidió reconocer las elecciones hondureñas y al

Government in exile Website

You are viewing an archived web page, collected at the request of University of Texas at Austin Libraries, Latin American Government Documents Archive using [Archive-It](#). This page was captured on 20:36:57 Sep 19, 2009, and is part of the [Latin American Government Documents Archive, LAGDA](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page.

República de **HONDURAS**

Manuel Zelaya Rosales **Presidente 2006-2010**


PRESIDENTE
Oficina de la PRIMERA DAMA
del GOBIERNO
MINISTRO de la PRESIDENCIA
CANCILLER
SECRETARIO Priv.

GOLPE de ESTADO

La madrugada del 28 de Junio del 2009 el Presidente Zelaya fué secuestrado y trasladado a Costa Rica por fuerzas militares, y desde esa fecha le impiden su retorno al País. El gobierno golpista se ha ganado el repudio mundial y el rechazo de un pueblo que hoy es víctima de la opresión y abuso de esos dictadores

ANTES DEL GOLPE

DESPUES DEL GOLPE

1

2

3

4



GOLPE DE ESTADO

GOBIERNO DEL PRESIDENTE JOSE MANUEL ZELAYA

INFORME PRELIMINAR SOBRE EL GOLPE DE ESTADO

I- CONSIDERACIONES GENERALES.

II- AUTORES DEL GOLPE DE ESTADO. CATEGORIAS.

a. Primera Categoría

Documentos


Preliminary Report on The COUP D'ETAT (Resumen) 10-Sept-2009


Report on The COUP D'ETAT (All) 10-Sept-2009


Resumen Ejecutivo: Informe Preliminar sobre el Golpe de Estado. Gobierno del Presidente Manuel Zelaya. 10-Sept-2009


Informe Completo sobre el Golpe de Estado. Gobierno del Presidente Manuel Zelaya 10-Sept-2009

Why Web Archive



Why Web Archive



Why Web Archive



WHAT STARTS HERE CHANGES THE WORLD
THE UNIVERSITY OF TEXAS AT AUSTIN

University of Texas Libraries

[Map & Floorplans](#) | [Hours](#) | [InterLibrary Services](#)
[Article Databases](#) | [Other Catalogs](#)

Library Catalog: Search for Books, Journals, Music and More

[Help](#) | [Ask a Librarian](#) | [Renew & My Account](#)

Keyword

Advanced Search

Title

Journal Title

Author

Author + Title

Subject Heading

Call Number

ISBN or Other Nos.

Course Reserves

Music Search

[Start Over](#)
[Save to My Lists](#)
[Save to Clipboard](#)
[MARC Display](#)
[Return to List](#)
[Modify Search](#)
[More Like This](#)
[Another Search](#)

(Search History)

☐ [Limit to available items](#)

Did you mean [memorial secretarial mexico?](#) [more »](#)

277 results found. Sorted by [relevance](#) | [date](#) | [title](#) .

Results page: [Previous](#) [Next](#)

Corporate author [Mexico. Secretaría de Hacienda y Crédito Público.](#)

Title **Memoria provisional presentada al soberano Congreso / por el Ministerio de Hacienda en 2 de junio de 1823.** [\[Bookmark Link\]](#)

Publication Information [México] : Impr. Nacional del Supremo Gobierno, [1823]

Location	Call No.	Current Status
Benson Collection LAC-ZZ Rare Books	HJ 15 A22 1823M	LIB USE ONLY
Benson Collection LAC-ZZ Rare Books	HJ 15 A22 1823M c.2	LIB USE ONLY

Description 16 p. ; 30 cm.

Note Signed: Francisco de Arrillaga.

Local note Copy 1 with: **Memoria** que el secretario de estado y del despacho de hacienda presentó al soberano Congreso Constituyente sobre los ramos del Ministerio de su cargo, leida en la sesión del día 12 de noviembre de 1823. México : Impre. del Supremo Gobierno, [1823]

Subject [Finance, Public](#) -- [Mexico.](#)

Added author [Arrillaga, Francisco de.](#)

OCLC number 9539252

History of archiving Latin America at UT Austin

- Benson Library collected gov docs in print since 1920s
- Latin America began moving to digital gov docs around 2000
 - Download, print and curate
- Latin American Government Document Archive begins 2005
 - Crawl entire websites, compress and curate data
 - Provide access to digital content directly



Latin American Web Archiving Project

[Home](#)
[LAGDA](#)
[México 2010](#)
[ARVEPODIS](#)
[APPELA](#)

LAGDA Links

[LAGDA Home](#)
[Browse All Government Documents](#)
[Sample Presidential Messages](#)
[Sample Ministerial Documents](#)

Conduct a full text search of LAGDA:

Spotlight



[Full text](#) of Honduran
President Manuel Zelaya's
speeches, 2006 - 2008

Latin American Government Documents Archive

[English](#)
[Español](#)
[Português](#)

About the Archive

The *Latin American Government Documents Archive* (LAGDA) seeks to preserve and facilitate access to a wide range of ministerial and presidential documents from 18 Latin American and Caribbean countries. The Archive contains copies of the Web sites of approximately 300 government ministries and presidencies. Capture of sites began on multiple dates in 2005 and 2006, and will continue with regularly scheduled captures.

Content in the Archive includes not only the full-text versions of official documents, but also original video and audio recordings of key regional leaders. Archive contents include thousands of annual and "state of the nation" reports; plans and programs; and speeches by presidents and government ministers. Content can be accessed via full-text search ([search help](#)), or by browsing by country or by specialized sample collection, such as "Presidential Messages" or "Ministerial Documents."

LAGDA is a joint project of the University of Texas Libraries, The Nettie Lee Benson Latin American Collection, and the Latin American Network Information Center at The University of Texas at Austin. Web archiving services are provided by the Internet Archive's [Archive-It](#) service.

If you know of a Latin American government Web site that you feel an archived version would be useful to your research and it is not currently on [our list](#), we invite you to [Nominate a Site for inclusion in the LAGDA Web archive](#).

Browse LAGDA

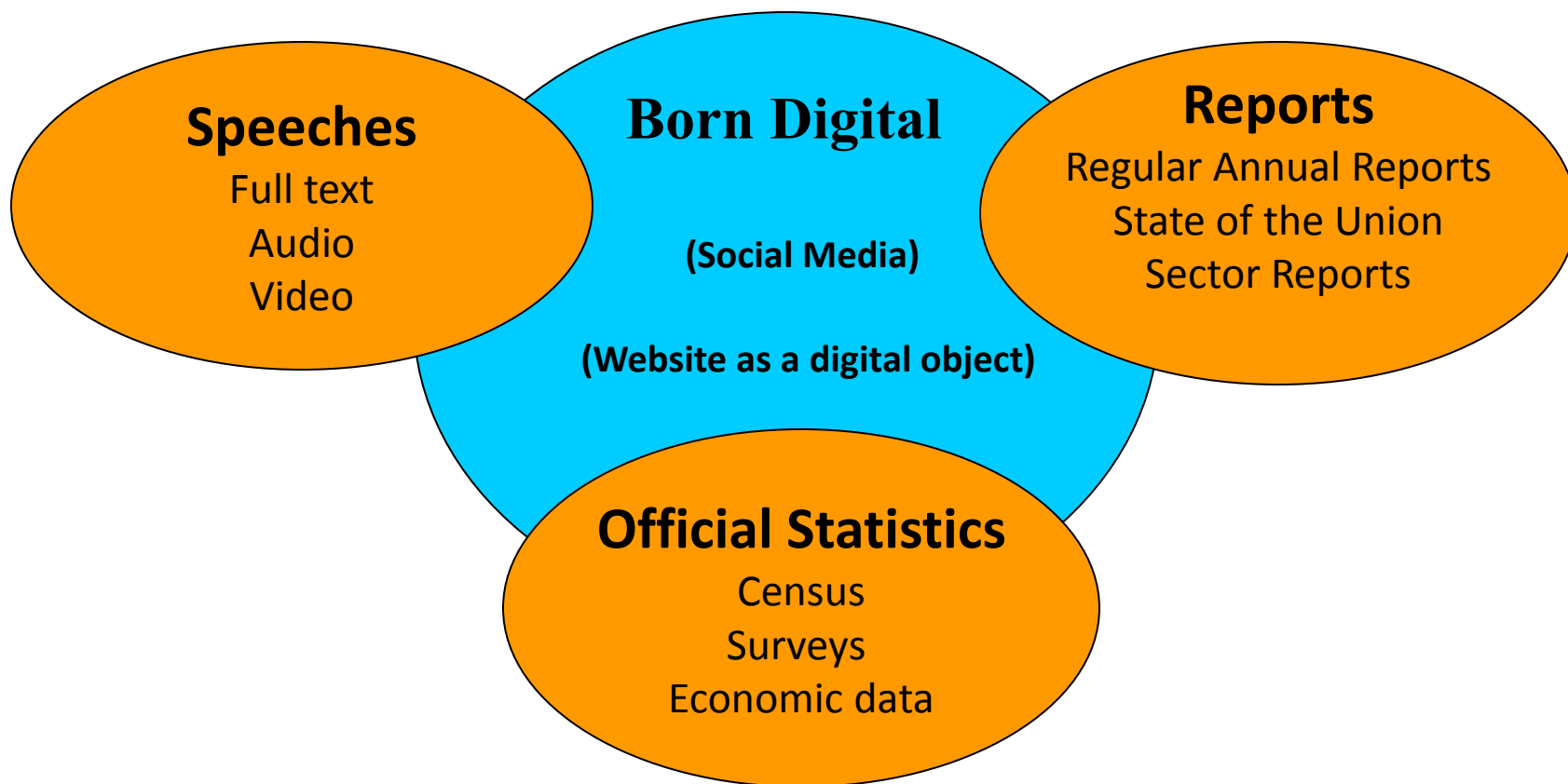
- ◆ [Full collection](#) (or full collection in [Proxy Mode](#))
- ◆ [Presidential Messages & Speeches](#)
- ◆ [Sample Ministerial Documents](#)

Latin American Government Document Archive

LAGDA = **280 seeds**, about 15 government ministries per each of 18 countries crawled **quarterly** since 2005

- Files crawled and archived to date in LAGDA **70 million**
- Data archived **5.9 TB**
- Items added to collection per year **9-10 million**
- HTML pages archived per crawl **1.6 million**
- PDF documents archived per crawl **260,000**
- Monthly average pageviews on LAGDA **2,918**

Latin American Government Documents



LAGDA: challenges to data mining

- Heterogeneous corpus
 - Various languages
 - Data formats (HTML, Word, PDF, Other)
 - Document characteristics
- Minimal metadata
- Variety of sources (countries, governments, departments)

LAGDA: motivating problem

- Goal:
 - Automatically attach labels to documents in a large collection based on training documents
- Challenges:
 - Keyword search is ineffective due to lack of consistent words
 - Training documents may cover broad subject areas

LAGDA: techniques for data mining

- Break documents into n-grams
 - 1-gram {The, quick, brown, fox, jumps, over, the, lazy}
 - 2-gram {The quick, quick brown, brown fox, fox jumps}
 - 3-gram {The quick brown, quick brown fox...}
- Identify one or more subsets of n-grams with significant high usages in the training documents
- Evaluate all documents in the corpus using these n-grams



LAGDA: techniques for data mining

- Use this score and others to create a composite score
- The company you keep - Examine the text and the links that point to our documents
- Natural language processing
 - Named entities & Part-of-Speech tagging

LAGDA: technology for large-scale computing at TACC

- Corral data storage system (6 Petabytes)
- Longhorn High Performance Cluster
- Paradigms for distributed computing (MPI and Hadoop)
 - Nodes work in parallel and combine their results
 - Allows us to divide and conquer the problem
- Open source libraries (Heritrix, Tika, Lucene, OpenNLP)



LAGDA: initial results

- Traditional classification approaches are unsuccessful
- Our n-gram approach for classification based on training set outperforms traditional Bayesian Inference Classifier
- Results from our composite scores demonstrate additional improvement

“big data” and libraries: going forward

- Challenges posed by web-archived data
 - Size, heterogeneity and limited metadata
- Data access that is more dynamic and flexible
- How big data can create data-driven research
- Development of use cases and research examples
- Technology at the service of social sciences, humanities and other fields whose research could benefit

Acknowledgments

- Kent Norsworthy, LLILAS and Benson Collection
- Weijia Xu, TACC
- Carolyn Palaima, LLILAS and Benson Collection
- UT Libraries

-  **TEXAS ADVANCED COMPUTING CENTER**
Powering Discoveries That Change The World



Contact

nwoodward@mail.utexas.edu

<http://lanic.utexas.edu/project/archives/lagda/>

<http://www.archive-it.org/public/collection.html?id=176>

Google: LAGDA