Data Management in the UNT Digital Library

Mark Phillips
UNT Libraries
October 1, 2014
UNT Digital Library
Serves as infrastructure and access for research datasets at the UNT Libraries
UNT Data Repository
What's Inside this Collection.

The UNT Data Repository is a central archive to provide long-term preservation and access to the research data of the UNT scholarly community. The Data Repository works in conjunction with the UNT Scholarly Works repository to ensure long-term access to the full range of research outputs from UNT.

Browse the holdings of this collection
Complements other repository collections
UNT Scholarly Works Repository
The Scholarly Works Collection is home to materials from the UNT community's research, creative, and scholarly activities. It serves as UNT's Open Access Repository. This collection brings together articles, papers, artwork, music, research data, reports, presentations, and other scholarly and creative products representing the expertise in our university community.

UNT Scholarly Works aims to:

- Provide easy access to valuable scholarly and creative materials from the UNT community
- Promote discovery through effective search and navigation tools
- Secure long-term access through stewardship and preservation
- Ensure sustainability through continuing system improvements
- Showcase UNT's research and creative achievements to a worldwide audience

To learn more about UNT Scholarly Works, see our web page or contact us at untrepository@unt.edu.
UNT Theses and Dissertations
Theses and dissertations represent a wealth of scholarly and artistic content created by masters and doctoral students in the degree-seeking process. In 1999, the University of North Texas was one of the first American universities to begin requiring electronic theses or dissertations (ETDs) for graduation. Some ETDs in this collection are restricted to use by the UNT community.
Allows for easy linkages (via metadata) to other items in the repository.
Dataset + Presentation
Dataset + Paper, Article, Report
Or just a dataset
Example Dataset in UNT Data Repository
Landing page for Dataset
Full Metadata Display
# Data Annex to the United Nations Truth Commission on the civil war in El Salvador from 1979--1991 (digitized text)

| Creator (Compiler):       | Mason, T. David  
University of North Texas Political Science |
|---------------------------|------------------------------------------------|
| Creator (Compiler):       | Hamner, Jesse  
Academic Computing and Technical Services,  
University of North Texas |
| Creator (Compiler):       | Phillips, Mark Edward  
University of North Texas Libraries |
| Date(s):                  |  Creation: October 2012 |
| Coverage:                 | Place  
El Salvador  
From 1979 to 1991 |
<p>| Description:              | This dataset contains statistical information transcribed from the supplementary documentation of a United Nations (UN) report compiled by The Commission on the Truth for El Salvador (La Comisión de la Verdad para El Salvador). It includes information about approximately 20,000 civilian/noncombatant victims of the civil war in El Salvador (from 1979 to 1991) taken from interviews of those who survived or knew/knew of those who were victims. |
| Note:                     | Both the original, raw data and modified files are available. For more information, see the &quot;readme.txt&quot; file. |
| Physical Description:     | 4 files: Anexos-Tomo-II-Informe-Comision-Verdad-para-El-Salvador.pdf (29 MB), el-salvador.csv (1.4 MB), el-salvador-master-file.csv (1.1 MB), el-salvador-geocoded-file.txt (1.6 MB) |</p>
<table>
<thead>
<tr>
<th>Note:</th>
<th>Both the original, raw data and modified files are available. For more information, see the &quot;readme.txt&quot; file.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Description:</td>
<td>4 files: Anexos-Tomo-II-Informe-Comision-Verdad-para-El-Salvador.pdf (29 MB), el-salvador.csv (1.4 MB), el-salvador-master-file.csv (1.1 MB), el-salvador-geocoded-file.txt (1.6 MB)</td>
</tr>
<tr>
<td>Language(s):</td>
<td>Spanish</td>
</tr>
</tbody>
</table>
| Subject(s): | LCSH: Disappeared persons -- El Salvador.  
LCSH: Torture -- El Salvador. |
| Keyword(s): | military atrocities | persecution | civil wars | political unrest | victims | statistics | killings | homicides | kidnappings | homocidios | desapariciones forzadas | torturas | malos tratos | lesiones grandes | secuestros extorsivos | violaciones sexual |
| Contributor(s): | Author: La Comision de la Verdad para El Salvador  
Other: United Nations Publisher of original text. |
| Partner: | UNT Libraries  
About | Browse this Partner |
| Collection: | UNT Data Repository  
About | Browse this Collection |
| Identifier: | ARK: ark:/67531/metado268920 |
| Resource Type: | Dataset |
| Format: | Text |
| Rights: | License: Public Domain |
Access to Dataset Files
Data Annex to the United Nations Truth Commission on the civil war in El Salvador from 1979--1991 (digitized text)

<table>
<thead>
<tr>
<th>Filename</th>
<th>Size</th>
<th>Format</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anexos-Tomo-II-Informa-Comision-Verdad-para-El-Salvador.pdf</td>
<td>27.5 MB</td>
<td>application/pdf</td>
<td>Download</td>
</tr>
<tr>
<td>README.txt</td>
<td>5.1 KB</td>
<td>text/plain</td>
<td>Download</td>
</tr>
<tr>
<td>el-salvador-geocoded-file.txt</td>
<td>1.5 MB</td>
<td>text/plain</td>
<td>Download</td>
</tr>
<tr>
<td>el-salvador-master-file.csv</td>
<td>1.1 MB</td>
<td>text/plain</td>
<td>Download</td>
</tr>
<tr>
<td>el-salvador.csv</td>
<td>1.3 MB</td>
<td>text/plain</td>
<td>Download</td>
</tr>
</tbody>
</table>

Feedback:
If you are having problems, need to report errors, or have questions or comments for the staff, please use our Feedback Form.
Dataset specific info is captured in README.txt
Summary:


The United Nations created a truth commission to uncover as many civilian/noncombatant victims of the civil war in El Salvador (from 1979 to 1991) as they could, using interviews of those who survived or knew/knew of those who were victims. The list contains approximately 20,000 individuals.

The digital files containing the list of individual victims and ancillary data are not at present available, or else have never been made available. This project transcribes and splits the paper files into digital records for use by others.

Institutional contact information:
untrepository@unt.edu

Additional information is available in the metadata files of the UNT data repository.

This set of files is version 1.0 of the data, October 2012 (uploaded to UNT in December 2013).

The original data set does not consistently include accents. These grammatical omissions from the original Spanish language data do not affect the fundamental characteristics of the events data.

Original titles and sections:

Anexos Tomo II: Informe Comision Verdad para El Salvador

6. Lista de victimas presentadas a la comision
   A. Fuente Directa
   victimas cuya identidad no es reservada
   B. Fuente Directa
   victimas cuya identidad es reservada
   C. Fuente Indirecta

Transcribed, error-checked and (where noted) geospatial data added by the Castleberry Peace Institute and The UNT Libraries, University of North Texas.

Authorship and copyright information:

The original data were produced by the United Nations. Those individuals who have transcribed and modified these data assert no claim to copyright.

Files:
Files:

Anexos Tomo II Informe Comision Verdad para El Salvador.pdf
- a PDF scan of the original data
- 29 MB

e-l-salvador.csv
- transcribed from the PDF
- known to have some errors, corrected at least in part
  in the "master" file.
- 1.4 MB

e-l-salvador-master-file.csv
- modified data from the transcribed file
- machine-checked for errors
- 1.1 MB

e-l-salvador-geocoded-file.txt
- tab-separated text
- same data as in the "master" file
- geospatial reference data added (but not long-lat coordinates)
- 1.6 MB

Data formats:

The raw (unedited) file has the following format:

APPELLIDOS: Surname
NOMBRES: Given name
HECHO: Method of injury
FECHA: Date of event in dd/mm/yy
LUGAR: Location code (three consecutive two- or three-digit locations)
Resp1: Abbreviation of the responsible party (primary)
Resp2: Abbreviation of an additional responsible party
Resp3: Abbreviation of an additional responsible party
Resp4: Abbreviation of an additional responsible party

The corrected file has the following format:

APPELLIDOS: Surname
NOMBRES: Given name
HECHO: Method of injury
FECHA: Date of event in dd/mm/yy
LUGAR: Location code (three consecutive two- or three-digit locations)
Resp1: Abbreviation of the responsible party (primary)
Resp2: Abbreviation of an additional responsible party
Resp3: Abbreviation of an additional responsible party
Resp4: Abbreviation of an additional responsible party
Error: Error caught by a human or a program (1=yes)
Column originated: column in which the error was found
Fixed: Was the error fixed? (1=yes)
Notes: Notes on the error

The geocoded (you should use this file) has the following format:

AbuseType: single-letter code for the event type
(If not listed or error (not present in the original file)
The raw (unedited) file has the following format:

APPELLIDOS: Surname
NOMBRES: Given name
HECHO: Method of injury
FECHA: Date of event in dd/mm/yy
LUGAR: Location code (three consecutive two- or three-digit locations)
Resp1: Abbreviation of the responsible party (primary)
Resp2: Abbreviation of an additional responsible party
Resp3: Abbreviation of an additional responsible party
Resp4: Abbreviation of an additional responsible party

The corrected file has the following format:

APPELLIDOS: Surname
NOMBRES: Given name
HECHO: Method of injury
FECHA: Date of event in dd/mm/yy
LUGAR: Location code (three consecutive two- or three-digit locations)
Resp1: Abbreviation of the responsible party (primary)
Resp2: Abbreviation of an additional responsible party
Resp3: Abbreviation of an additional responsible party
Resp4: Abbreviation of an additional responsible party
Error: Error caught by a human or a program (1=yes)
Column originated: column in which the error was found
Fixed: Was the error fixed? (1=yes)
Notes: Notes on the error

The geocoded (you should use this file) has the following format:

Abusstype: single-letter code for the event type
0: not listed or error (not present in the original file)
d: disappearance
h: homicide
l: wounds (lesiones)
o: other [methods]
s: kidnapping (secuestro)
t: torture
v: rape or sexual assault (violacion)

Depnum: Numeric code for Department
Dep: name of Department
Muninum: number of Municipio
Muni: name of Municipio
Cantnum: number code for
LOCOCODE: UN code for the data
FTPSDCODE: FTPS code for Salvadorean Department (HS-nn)
ISO3166CODE: ISO-3166 alphanumeric code for Department (SV-ll)
SLVNCODE: simple number code for Department
SLVMCODE: simple three-digit number code for Municipio
Day: Two digit day of the event, if it exists (99 otherwise)
Month: Two digit month of the event, if it exists (-99 otherwise)
Year: Four digit year, if it exists
Resp1: Abbreviation of the responsible party (primary)
Resp2: Abbreviation of an additional responsible party
Resp3: Abbreviation of an additional responsible party
Resp4: Abbreviation of an additional responsible party

-- END OF FILE --
User can download dataset
Example of another README.txt
Readme for YesAllWomen Tweets (JSON) Dataset:

This dataset is being made available using a Creative Commons CC0 Public Domain Dedication. See http://creativecommons.org/publicdomain/zero/1.0/

Data contained within each tweet sample is copyright by the tweet's authors and is subject to all copyright law. The creator of the dataset makes no assertion of rights related to the content of each tweet.

The dataset is provided 'as is' without warranty or any representation of accuracy, timeliness or completeness.

The dataset format is line-oriented JSON in utf8 encoding.

Dataset Overview:

This dataset contains Twitter JSON data for several Twitter search queries that were collected related to the "Yes All Women" conversation on Twitter using the twarc (https://github.com/edub/twarc) package that makes use of Twitter's search API. The following searches are included in this dataset.

#YesAllWomen
#NotAllMen
#NotAllWomen

Twitter data making up this dataset was captured between 2014-05-25 and 2014-06-08.

All Twitter data is stored as line-oriented JSON (each line is a complete JSON document), and is exactly what was received from the Twitter API.

Public Files:

NotAllMen.ids : 69,873 Unique Twitter IDs for the search #NotAllMen
NotAllMen_images.tsv : 7,426 Unique Twitter IDs and mapping to image URLs for #NotAllMen
NotAllWomen.ids : 1,827 Unique Twitter IDs for the search #NotAllWomen
NotAllWomen_images.tsv : 103 Unique Twitter IDs and mapping to image URLs for #NotAllWomen
YesAllWomen.ids : 2,705,985 Unique Twitter IDs for the search #YesAllWomen
YesAllWomen_images.tsv : 714,459 Unique Twitter IDs and mapping to image URLs for #YesAllWomen

Restricted Files:

images.tar : 34,532 Images harvested from #NotAllMen #NotAllWomen and #YesAllWomen Tweets
NotAllMen.json.gz : 70,569 Tweets for search #NotAllMen
NotAllWomen.json.gz : 1,836 Tweets for search #NotAllWomen
YesAllWomen.json.gz : 2,733,358 Tweets for search #YesAllWomen

If you find errors or have questions, contact Mark Phillips at mark.phillips@unt.edu
Metadata is edited with standard metadata editing system.
Title

Select a qualifier for this title

Main Title

Enter a value for this title

The 0.22 Percent Across-the-Board Cut in FY2001 Appropriations

Creator

The name of a person, organization, or event (conference, meeting, etc.) associated in some way with the resource

Keith, Robert

Used to designate if a creator is an individual or an organization

Personal

The role that the person or organization played in the creation of the item

Author

Additional information about the creator related to the specific item

Fields

The name of a person, organization, or event (conference, meeting, etc.) associated in some way with the resource

Government and Finance Division

Used to designate if a creator is an individual or an organization

Organization

The role that the person or organization played in the creation of the item
Most of the work with depositors is in the area of file formats and creating the readme.txt files.
They generally seem willing to create readme.txt files and seem to understand their importance.
We provide guidance and a template for creating the readme.txt files
Other descriptive metadata is generally standard descriptive metadata.
UNT Data Days
Day long data events that allow faculty to work with digital library personnel to get over the hurdles of depositing data.
Questions that are still open:
Scope of collect:
UNT Data only?
What about data we collect as a library, should it go in there are well?
We are thinking yes.
Questions?

http://digital.library.unt.edu/datarepository