<https://web.archive.org/web/20141224002711/http://watchdoglabs.org/blog/2012/12/14/cleaning-and-standardizing-white-house-visitor-data>

## [Cleaning and standardizing White House visitor data](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/2012/12/14/cleaning-and-standardizing-white-house-visitor-data/)

[*efg*](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/author/efg/)[*December 14, 2012 at 12:11 am*](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/2012/12/)

By Earl F Glynn | Franklin Center

This is the fourth technical article in a series about analyzing White House visitor data. This article looks at cleaning and standardizing visitor data, and creating descriptive statistics of the data fields.



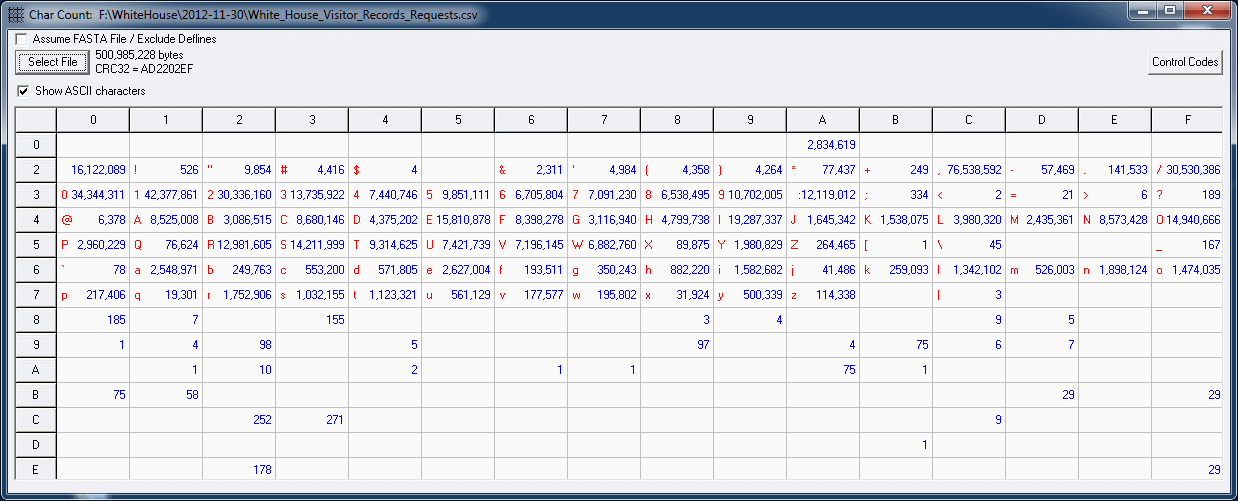
The complete White House visitor data file released each month has a number of data problems that need to be cleaned up before working with the file.

Problems are encountered in most of the data fields of interest, including names, dates, and locations.  Standardization of formats and values are introduced for some of the key fields to simplify searches and analysis.

See the many technical details of the cleanup and standardization below.

**“Problem” characters in file?**

An [earlier article explained how to download the complete visitor data file via the Socrata data player](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/2012/12/11/discrepancies-between-raw-format-and-final-release-versions-of-white-house-visitor-records/):  **White\_House\_Visitor\_Records\_Requests.csv.**

[](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/wp-content/uploads/2012/12/White-House-Visitor-Records-Requests-CharCount-2012-11-30.gif)

*Output from CharCount utility showing frequency counts of all ASCII characters in White House visitor file from Nov 30, 2012. (Click on graphic to enlarge)*

Having seen a number of problem data files over the years, I often use a [CharCount utility](https://web.archive.org/web/20141224002711/http:/www.efg2.com/Lab/OtherProjects/index.html" \o "CharCount Utility" \t "_blank) to inspect a file for  “problem” characters.

[ASCII codes](https://web.archive.org/web/20141224002711/http:/www.ascii-code.com/) are fairly standard for 128 characters (x’00′ to x’7F’), but the 128 “extended ASCII codes” from x’80′ to x’FF’ are not.

Output from the CharCount utility (click on graphic at right) shows a small number of characters in the “extended” range.  Some of these characters are interpreted differently on [Macintosh](https://web.archive.org/web/20141224002711/http:/academic.evergreen.edu/projects/biophysics/technotes/program/ascii_ext-mac.htm), Linux and [Windows](https://web.archive.org/web/20141224002711/http:/www.ascii-code.com/) computers.

Since the July 27, 2012 released visitor file, three “new” characters now appear in the file:  x’92′, x’A2′, x’B1′.  Macs and Windows computers will interpret x’92′ differently.  The characters x’A2′ (¢) and x’B1′ (±) appear to be interpreted the same way, but it’s unclear why they are now in the file.

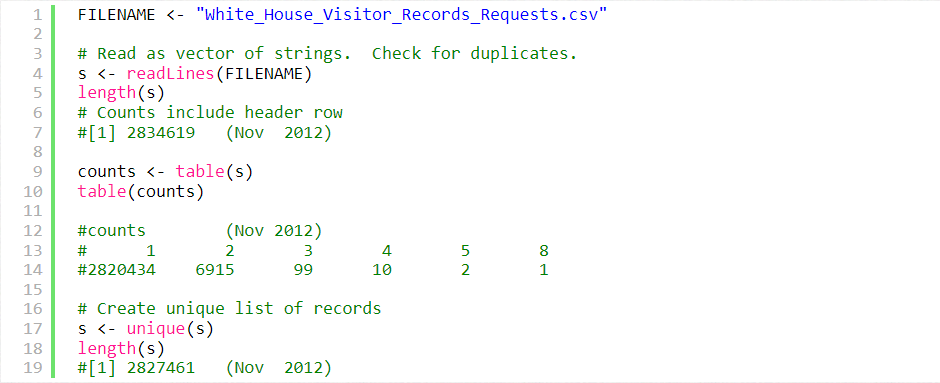
The “problem” characters could be removed or replaced with something like the [Linux sed command](https://web.archive.org/web/20141224002711/http:/www.linuxjournal.com/article/7231), but for now these problem characters will be ignored.

Some names with non-English characters might include these special characters and could be difficult to find, or might sort in an unexpected order.

**Duplicate Records**

I wrote the **[WhiteHouse-Visitors.R](https://web.archive.org/web/20141224002711/http:/www.watchdoglabs.org/data/WhiteHouse/2012-11-30/WhiteHouse-Visitors.R" \o "WhiteHouse-Visitors.R" \t "_blank)** script, and refined it over the past few months, to define a repeatable process for working with the White House visitor data.

This script first read the text file as a vector of strings, and looked for duplicates using R’s table function  (comments in green show results):

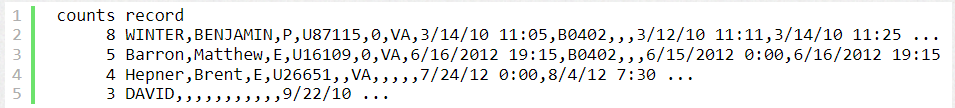


The counts vector above shows 1 record occurred 8 times, 2 records occurred 5 times, 10 records occurred 4 times, 99 records occurred 3 times, and 6915 records occurred twice.

The remaining 2,820,434 records were unique and occurred only once.

The file [**WhiteHouse-Visitors-Duplicate-Records.txt**](https://web.archive.org/web/20141224002711/http:/www.watchdoglabs.org/data/WhiteHouse/2012-11-30/Problems/WhiteHouse-Visitors-Duplicate-Records.txt) shows the duplicates.

Selected duplicates are shown below:



It’s fairly unlikely that 8 different people named “Benjamin P Winter” all attended the same White House meeting on March 14, 2010.  Likewise, it’s fairly unlikely that three different people named “Matthew E Barron” or four people named “Brent E Hepner” attended the same White House meetings.  The cause of such duplicates is unknown, but could be human error.

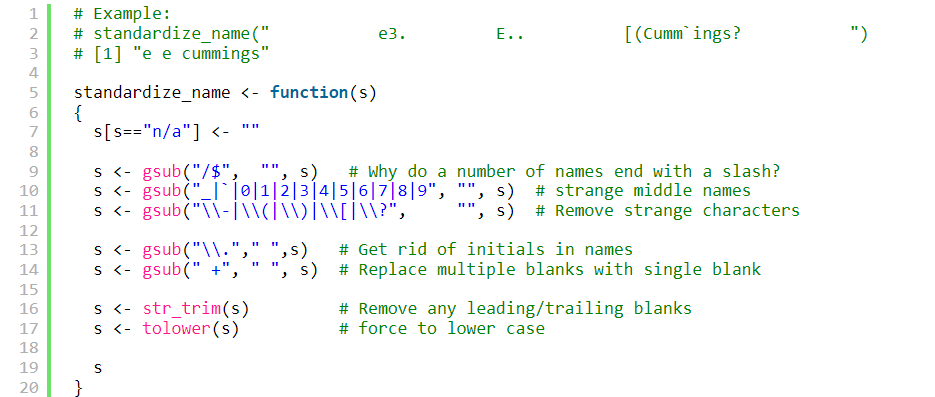
There are several cases in the file of apparent duplicates visitors attending a Rosh Hashana Reception at the Vice President’s Residence (code VPR) where only a single first name is given as the last name.  In this case, it’s quite plausible three different people named “David” attended the event, but without any other information, there is no information loss caused by deletion of the duplicates.

When all the duplicate strings were removed using R’s unique function, the file was found to have **2,827,461 unique records** (including the header record).

**Standardizing all names**

Inspection of various name fields (visitor, visitee, caller) shows a number of problems.  For example, names should not contain numbers (e.g., 2, 5, 9) or punctuation (e.g., [ or ?).

The R standardize\_name function (in separate file **[WhiteHouse-SharedFunctions.R](https://web.archive.org/web/20141224002711/http:/www.watchdoglabs.org/data/WhiteHouse/2012-11-30/WhiteHouse-SharedFunctions.R" \t "_blank)**) was written to make names more consistent so searches and sort orders would be more predictable:



Hyphenated last names sometimes have hyphens and sometimes do not. Perhaps hyphens should be retained in such last names, but for uniformity and consistency in name searches, all hyphens were removed.

**Removing records with missing visitor names**

If the point of the visitor records is to identify a visitor by name, it doesn't make sense to keep records without complete names -- at least a first and a last name.

The file [**WhiteHouse-Visitors-Missing-Names.txt**](https://web.archive.org/web/20141224002711/http:/www.watchdoglabs.org/data/WhiteHouse/2012-11-30/Problems/WhiteHouse-Visitors-Missing-Names.txt) (tab delimited file) shows 326 records with incomplete names.

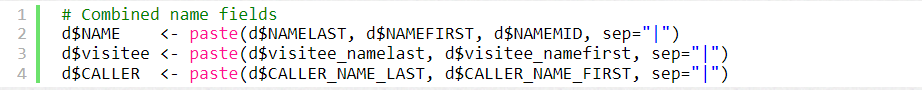
One record from the 3/30/2012 release is missing both names and all other information.

The majority of the records with incomplete names are from visits to the Vice President's Residence (VPR) released on 12/31/2010 where only a first name (in the last name column) was specified.

**Combined name fields**

In the original file the visitor names are broken into three fields (NAMELAST, NAMEFIRST, NAMEMID), while the visitee and caller names are broken into two fields (e.g., visitee\_namelast and visitee\_namefirst).

To enable searching for a name  in a single field, combined names were introduced using a pipe ("|") separator:  NAME (for visitor), visitee and CALLER.



Examples:

NAME:

* dunn|anita|b
* geithner|timothy|f
* reich|robert|b

visitee:

* axelrod|david
* office|visitors        [tourists]
* plouffe|david
* potus|                [President of the United States]

CALLER:

* doebler|max
* gallagher|clare

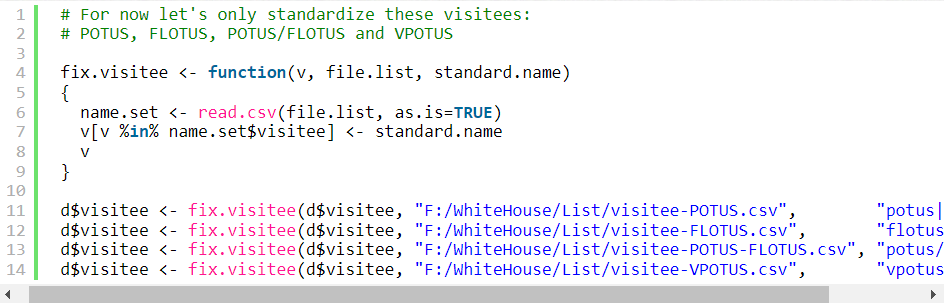
**Standardizing selected visitee names**

Unfortunately, there is little standardization of names in the file.  Names of the same person on different visits can be specified in various ways.

For now it does not seem wise to introduce any standardization of visitor names, but at least certain visitee names need standardization.

With the visitee names, I have identified [100 variations of “POTUS”](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/data/WhiteHouse/List/visitee-POTUS.csv) (President of the United States), [10 variations of “FLOTUS”](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/data/WhiteHouse/List/visitee-FLOTUS.csv) (First Lady of the United States), [5 variations of POTUS/FLOTUS](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/data/WhiteHouse/List/visitee-POTUS-FLOTUS.csv) combinations, and [6 variations of VPOTUS](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/data/WhiteHouse/List/visitee-VPOTUS.csv) (Vice President of the United States).  Other variations may be found in the future.

I combine all known variations for a single person in a file now, and use a simple R function to assign the same standard name to all the variations:



Additional standardization of visitee names may be introduced after studying known White House staff names.

**Single location field and standardization**

Similar to the treatment of names, a single LOCATION field was introduced to be the combination of MEETING\_LOC (building) and MEETING\_ROOM.

Examples:

* wh|oval office
* wh|rose garden
* wh|situation room
* oeob|bowling alley

The visitor file lacks standardization of location names, e.g., 21 different LOCATION values have been identified to mean “wh|oval office”, 6 variations identify the “wh|rose garden”, 27 variations identify the “wh|situation room”, and 39 variations identify the “oeob|bowling alley”.

At present the R script imposes some standardization for 18 different locations in a way similar to the standardization of visitee values.  (See[“location” files in this directory](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/data/WhiteHouse/List/).)

Additional standardization may be added for selected locations at a later date.

**Standardization of date/time fields**

The White House visitor data has five date/time fields:

* APPT\_MADE\_DATE,
* APPT\_START\_DATE,
* APPT\_CANCEL\_DATE,
* LastEntryDate, and
* RELEASE\_DATE.

(See [White House WAVES Data Fields](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/2012/12/12/white-house-worker-and-visitor-entry-system-data-fields/) for more information).

Date/time fields were observed to have a number of inconsistencies:

* Extra leading or trailing blanks.
* Most fields had dates and times, but some only had dates.
* Some times had a 12-hour clock, others had a 24-hour clock.
* Some 12-hour clock dates used am/pm, other used AM/PM.
* Some dates had 2-digit years, others had 4-digit years.
* Some dates mysteriously were numbers like 40619.61714.

After some programmatic editing, all dates fields were convert to an ISO 8601 format (yyyy-mm-dd HH:MM:SS) using R’s strptime function.  See the function fix.WhiteHouse.dates in the file **[WhiteHouse-SharedFunctions.R](https://web.archive.org/web/20141224002711/http:/www.watchdoglabs.org/data/WhiteHouse/2012-11-30/WhiteHouse-SharedFunctions.R" \t "_blank)**.

The ISO 8601 format allows dates as character strings to be sorted into chronological order.

The most unexpected conversions needed were the numbers, like 40619.61714.

These numbers can be entered into Excel and formatted as dates to reveal what that really mean, such as 3/17/2011 2:48 PM for 40619.61714.  This [online tip was found to do this conversion in R](https://web.archive.org/web/20141224002711/http:/r.789695.n4.nabble.com/Convert-number-to-Date-td1691251.html#a1691276).

A few date/time values were problems that have not been fixed programmatically yet:

* 6/1/20091 1:39:23AM
* 8/14/20091 0:29:51AM
* 8/31/200 912:00:00PM
* 9/18/200 912:01:00PM

**Standardization of the description field**

The description fields sometime contains comments about a visit, but text was entered with a number of inconsistencies.

Here are programmatic changes made to the description fields to make them more consistent:

* All letters were forced to lower case.
* Leading and trailing blanks were removed.
* Multiple internal spaces were replaced by a single space.
* Slashes or periods at the end of description fields were removed.

**“Cleaned-Enhanced” Files**

After making all the changes and standardizations described above to the White House visitor file in R, [“cleaned enhanced files” were created in this online “CLEANED-ENHANCED” directory](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/data/WhiteHouse/CLEANED-ENHANCED/) as a new starting point for any analysis.

These files are all tab-delimited.

The 715 MB file **WhiteHouse-Visitor-Records.txt** has 2,827,135 visit records.  [2,827,461 unique visit records – 326 records missing names).

The 715 MB file was zipped into a 121 MB file, [WhiteHouse-Visitor-Records-all-through-2012-11-30-Release.ZIP](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/data/WhiteHouse/CLEANED-ENHANCED/WhiteHouse-Visitor-Records-all-through-2012-11-30-Release.ZIP) for download.

Separate monthly files for all release dates are also available in the CLEANED-ENHANCED directory.

The new 82,614 visit records released on Nov. 30, 2012 can be downloaded in a separate 21 MB file:

* [**WhiteHouse-Visitor-Records-2012-11-30.txt**](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/data/WhiteHouse/CLEANED-ENHANCED/WhiteHouse-Visitor-Records-2012-11-30.txt)

**Metadata**

Frequency counts for all values in each data field (including the added NAME, visitee, CALLER and LOCATION fields) were created for the full visitor file, as well as a separate file with the statistics only for the Nov. 30, 2012 release.

Each file in the following directories shows the frequency counts for all possible values found:

* [Counts-ALL Directory](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/data/WhiteHouse/2012-11-30/Counts-ALL/) (counts for all released visitor records)
* [Counts-RELEASE Directory](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/data/WhiteHouse/2012-11-30/Counts-RELEASE/) (counts for Nov. 30, 2012 newly released data)

In addition to the single files by data field, a summary file of all fields was created corresponding to the ALL and RELEASE sets:

* [WhiteHouse-Visitor-Log-Field-Summary-ALL.csv](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/data/WhiteHouse/2012-11-30/WhiteHouse-Visitor-Log-Field-Summary-ALL.csv)
* [WhiteHouse-Visitor-Log-Field-Summary-RELEASE.csv](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/data/WhiteHouse/2012-11-30/WhiteHouse-Visitor-Log-Field-Summary-RELEASE.csv)

See Table 3 in [White House WAVES Data Fields](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/2012/12/12/white-house-worker-and-visitor-entry-system-data-fields/) for a formatted version of the ALL summary file.

Next article:  The cleaned-enhanced file  ***WhiteHouse-Visitor-Records.txt*** will be divided into categories to simplify analysis of POTUS and White House Staff visits.

**Series about analyzing White House visitor data**

* [Background information on White House visitor data](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/2012/12/10/background-information-on-white-house-visitor-data/), WatchdogLabs.org, Dec. 10, 2012.
* [Discrepancies between “raw format” and final release versions of White House visitor records](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/2012/12/11/discrepancies-between-raw-format-and-final-release-versions-of-white-house-visitor-records/), WatchdogLabs.org, Dec. 11, 2012.
* [White House Worker and Visitor Entry System Data Fields](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/2012/12/12/white-house-worker-and-visitor-entry-system-data-fields/), WatchdogLabs.org, Dec. 12, 2012.
* **Cleaning and standardizing White House visitor data, WatchdogLabs.org, Dec. 14, 2012.**
* [Two-thirds of White House visitor records about tourists](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/category/white-house/), WatchdogLabs.org, March 6, 2013.
* [Charts show end of White House tourist visits](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/2013/08/01/charts-show-end-of-white-house-tourist-visits/), WatchdogLabs.org, Aug. 1, 2013.
* [Out-of-cycle release of White House visitor records](https://web.archive.org/web/20141224002711/http:/watchdoglabs.org/blog/2013/08/02/latest-white-house-visitor-records-show-13-more-shulman-visits-in-2009/), WatchdogLabs.org, Aug. 2, 2013.