https://web.archive.org/web/20141222165725/http://watchdoglabs.org/blog/2012/12/11/discrepancies-between-raw-format-and-final-release-versions-of-white-house-visitor-records/

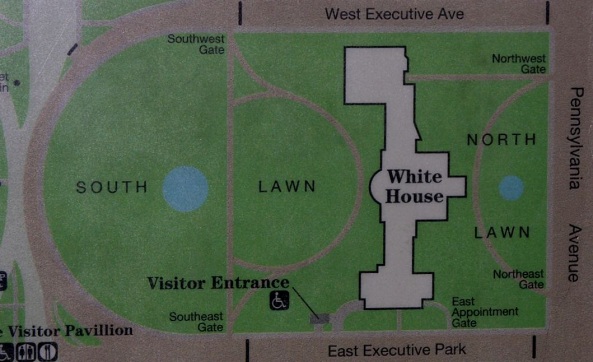


[Discrepancies between “raw format” and final release versions of White House visitor records](https://web.archive.org/web/20141222165725/http:/watchdoglabs.org/blog/2012/12/11/discrepancies-between-raw-format-and-final-release-versions-of-white-house-visitor-records/)

[*efg*](https://web.archive.org/web/20141222165725/http:/watchdoglabs.org/blog/author/efg/)[*December 11, 2012 at 9:25 pm*](https://web.archive.org/web/20141222165725/http:/watchdoglabs.org/blog/2012/12/)

By Earl F Glynn | Franklin Center

*This is the second technical article in a series about analyzing White House visitor data. This article looks at the differences between two versions of the released visitor data.*



Since December 2009 [White House Visitor Access](https://web.archive.org/web/20141222165725/http:/www.whitehouse.gov/briefing-room/disclosures/visitor-records) data has been updated on the last Friday of every month.  The most recent update on November 30 announced [More Than 2.8 Million Records Released](https://web.archive.org/web/20141222165725/http:/www.whitehouse.gov/briefing-room/disclosures/visitor-records).

With each monthly release there are two versions of what should be the same data, the single complete data file and a set of “raw format” files.  What are the differences between these two versions?

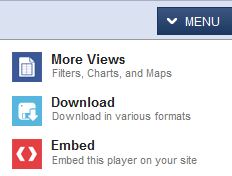
Analysis of the files indicates over 14,000 more records have been released in the single complete file than were released in raw format files.  These 14,000 records contained information about nearly 4,500 additional visitors seeing about 330 additional staff members at the White House complex.

The cause of the discrepancies between the two files that should have the same information is not known.

See analysis below.

**Complete visitor data file via Socrata data player**

The complete visitor data file can be downloaded from the [White House Visitor Access Records page](https://web.archive.org/web/20141222165725/http:/www.whitehouse.gov/briefing-room/disclosures/visitor-records)  in several formats via the “Menu” from the [Socrata Open Data Player](https://web.archive.org/web/20141222165725/http:/www.socrata.com/solutions/socrata-open-data-player/) hosted at [https://explore.data.gov](https://web.archive.org/web/20141222165725/https:/explore.data.gov/dataset/White-House-Visitor-Records-Requests/644b-gaut) .



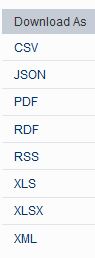
*Socrata data player “Menu” on White House Visitor Access Records page gives options to download the complete file, or embed the data player on another web page (see below).*

Using the HTML embed codes from the “Embed” Menu option shown above, the White House Visitor Access Records hosted at data.gov can be displayed on any web site, such as the live data player below:



[Powered by Socrata](https://web.archive.org/web/20141222165725/http:/www.socrata.com/)

The Menu option provides several choices for downloading the complete file of White House visitor data:

[](https://web.archive.org/web/20141222165725/http:/watchdoglabs.org/blog/wp-content/uploads/2012/12/Socrata-WH-Menu-Download.jpg)

*Socrata data player menu options for downloading White House visitor records in various file formats.*

Those familiar with Excel XLS or XLSX formats should **NOT** pick those options for downloading the data.  XLS files are limited to 65,536 records and XLSX files are limited to 1,048,576 records.  The complete White House visitor data file is approaching 3 million records and gets larger every month.  It’s unclear why options are available that will not work with the data.

Many data analysis programs work easily with the [CSV (comma-separated-values) format](https://web.archive.org/web/20141222165725/http:/en.wikipedia.org/wiki/Comma-separated_values), so that is the format I recommend.

The current downloaded CSV file, **White\_House\_Visitor\_Records\_Requests.csv**, with **2,834,619 record**s (including header) has a size of **477 MB**.

There  is no option to download a CSV inside a ZIP, which would be much smaller — something like 60 MB instead of 477 MB.  As the White House data file grows, its size is becoming a problem for quick downloads.

XML and JSON options are not practical in my opinion because of the “bloat” in these file formats.  Uncompressed versions of these files are likely over 1 GB in size.

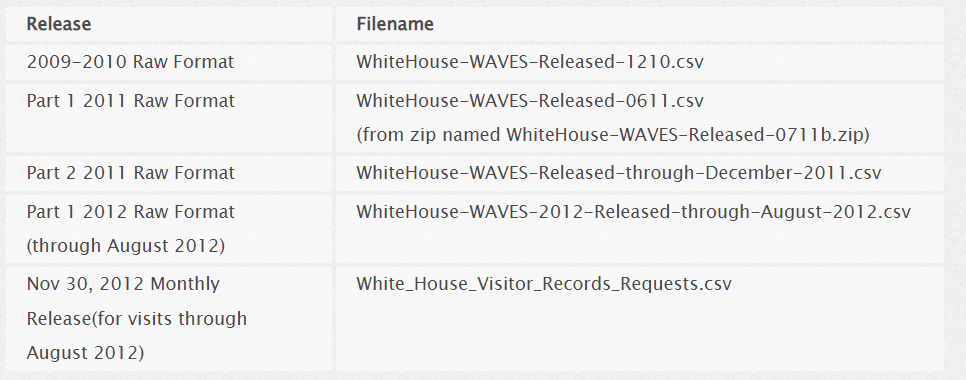
**“Raw Format” data**

Below the data player on the White House [Visitor Access Records](https://web.archive.org/web/20141222165725/http:/www.whitehouse.gov/briefing-room/disclosures/visitor-records) page there are a number of links to the “raw format” data:

*To download data released from 2009-2010 in its raw format,*[*click here*](https://web.archive.org/web/20141222165725/http:/www.whitehouse.gov/files/disclosures/visitors/WhiteHouse-WAVES-Released-1210.zip)*. (.zip of a .csv, 20MB)*  
*To download Part 1 of the data released in 2011 in its raw format,*[*click here*](https://web.archive.org/web/20141222165725/http:/www.whitehouse.gov/files/disclosures/visitors/WhiteHouse-WAVES-Released-0711b.zip)*. (.zip of a .csv, 7.4MB)*  
*To download Part 2 of the data released in 2011 in its raw format,*[*click here*](https://web.archive.org/web/20141222165725/http:/www.whitehouse.gov/files/disclosures/visitors/WhiteHouse-WAVES-Released-through-December-2011.zip)*. (.zip of a .csv, 15.2MB)*  
*To download Part 1 of the data released in 2012 (through August 2012) in its raw format,*[*click here*](https://web.archive.org/web/20141222165725/http:/whitehouse.gov/sites/default/files/visitors/whitehouse-waves-2012-released-through-august-2012.zip)*. (zip of a .csv, 17.7MB)*  
*To download an explanation of the column headers contained in the raw data file,*[*click here*](https://web.archive.org/web/20141222165725/http:/www.whitehouse.gov/files/disclosures/visitors/WhiteHouse-WAVES-Key-1209.txt)*. (.txt, 1.3KB)*

The following table shows the names of each of the raw format files downloaded from these links (with one exception, all CSV files have the same filenames as the ZIPs):

**Table 1.  Filenames of “Raw Format” White House visitor files**



I have not found an explanation for what processing was performed on the “raw format” files when they were combined to form the single final release file used with Socrata.

I wondered if some of the raw files might have some data not in other raw files, or not in the final combined file.  I wrote an [R script to study the files](https://web.archive.org/web/20141222165725/http:/watchdoglabs.org/data/WhiteHouse/2012-11-30/WAVES/WhiteHouse-WAVES.R) and find out.

**Comparing versions**

**Problems reading files**

The combined visitor file, **White\_House\_Visitor\_Records\_Requests.csv**, has 28 columns of data.

Raw format files were expected to have the same 28 columns of data, and sometimes they did.

The R script shows the details, but there were several problems in reading two of the raw format files.

All of the records in the file **WhiteHouse-WAVES-Released-0611.csv**had 256 comma separators for some reason, which means R saw 257 columns of data.  Most of these extra columns were “empty” fields and were simply deleted, but there was a problem with some of the records having data shifted into a 29th column.

The exact cause of the shifted data was not investigated.  The shifted data in column 29 were all Release.Date fields for these release dates:  4/27/2011, 5/27/2011, 6/24/2011.  For now the data shifted into column 29 were ignored.

The records in file **WhiteHouse-WAVES-Released-through-December-2011.csv** were a mess!  Roughly one-third of the records had the expected 28 data fields, but two-thirds had a number of “empty” fields caused by a variable number of commas at the end of many records.  Many records had 29 to 65 comma separators — many more than was expected.

For example, 96,390 records had 43 comma delimiters but all the extra commas were at the end of the records.  The R *gsub* function was used with a regular expression to delete all the extra commas following the Release.Date field.

Three of the records consisted of only commas with no data and were deleted.

I discovered 89,098 records had Release.Date information shifted from column 28 to column 29 because of an extra comma before the Release.Date field.  After removal of this extra comma, all of the Release.Date fields aligned in column 28.  [Something similar is likely the problem with the 0611.csv file, but that was never investigated.]

The December-2011.csv file appears to have been a combination of five other files and inadvertently included duplicate headers from those five files.  Four of the extra header rows were deleted programmatically.  One of the extra headers had been accidentally combined with another record, so its removal caused one data record to also be deleted.

After a number of programmatic changes  (see the R script), the December-2011.csv file finally had the expected 28 columns.  It’s entirely unclear what caused the numerous problems with this file and why this file was so inconsistent with all other flies.

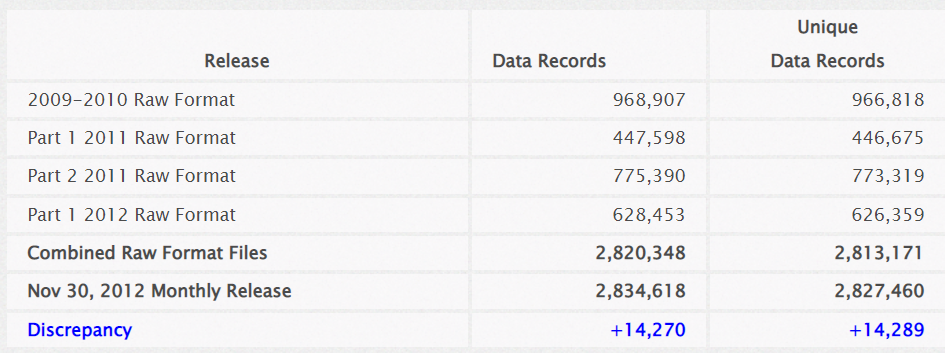
**Duplicate records**

Once all the files could be parsed into 28 fields, the R script used the *rbind* function to combine all the raw format files into a single data.frame in-memory for comparison with the single released file.

R provides a handy *unique* function that eliminates duplicate records (even with a data.frame), so the removal of duplicates was easy.

Table 2 shows the original number of records and the number of unique records in each of the files.

**Table 2.  Record counts in White House visitor data files**(excluding header record)



Overall, both the combined raw format files and the single released file have over 7,100 duplicates.

**Why so many duplicates?**

In theory, different people with exactly the same name could attend the same meeting and appear in the files as duplicates.

One would expect such duplicate problems to occur most likely with common names, like “John Smith”, but that has not been observed.

A future article will show additional information about duplicate visitor records, but I’ll give two examples here:  For some reason the visit by “Benjamin P Winter” on March 14, 2010 appears with 7 duplicates (the original record and 7 exact copies).   A June 15, 2012 meeting by “Matthew E Barron” had 4 duplicates (the original record and 4 exact copies).

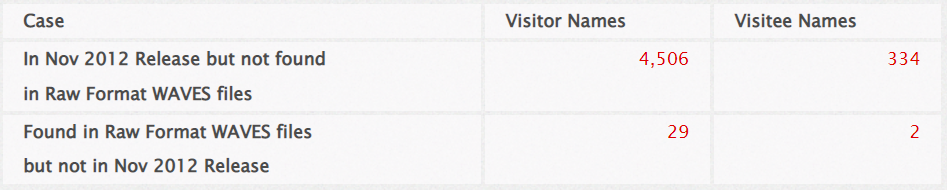
The cause for the duplicates is still not understood, except perhaps for human error.  **Can readers suggest other explanations for the duplicate records?**

**Discrepancies**

Table 2 shows the monthly single-file release has almost 14, 300 more records than can be found in the raw format files.

These 14,300 records resulted in a number of additional visitor and visitee names not released in the raw format files, as summarized in Table 3.

**Table 3.  Visitor/Visitee Name Discrepancies between White House Visitor Files**  
(click on link on number in table to see list of names)



**Names in Nov 2012 release but not found in raw format files:  4506 visitors, 334 visitees.**   The source of these “extra” names is not clear, but suggests there are some raw format records missing.

**Names found in raw format files but not in Nov 2012 release:  29 visitors, 2 visitees.**  Over half of the “extra” visitor last names were in the range “Zu” to “Zy” alphabetically.  This suggests a file sorted by last names may have been truncated in processing.

**Possible explanation for discrepancies?**

Why were more visitor records released in the single file than what can be found in the raw format files?

One possible explanation can be found in the [Declaration of Philp C. Droege [PDF page 54]](https://web.archive.org/web/20141222165725/http:/www.watchdoglabs.org/data/WhiteHouse/Legal/1-09-cv-02312-HHK-Jucicial-Watch-v-Secret-Service-13-Memorandum-of-Law.pdf), the White Office Office of Records Management Director, when he said on April 10, 2010:

*Since the announcement of the voluntary disclosure policy [September 4, 2009], the White House has made publicly available more than 2,500 WAVES visitor records created between January 20, 2009 and September 15, 2009, in response to individual requests from members of the public.*

With the White House responding to ongoing [“reasonable, narrow, and specific” requests for visitor records between January 20, 2009 and September 15, 2009](https://web.archive.org/web/20141222165725/http:/www.whitehouse.gov/requestvisitorrecords) (the subject of a [Judicial Watch law suit](https://web.archive.org/web/20141222165725/http:/www.watchdoglabs.org/data/WhiteHouse/Legal/1-09-cv-02312-BAH-Judicial-Watch-v-Secret-Service-01-Complaint.pdf)), perhaps additional records released as responses to those requests were added to the single file, but not the raw format files?  One problem with this theory is that I can find no evidence such records have been released by the White House in the last six months.

**Can readers suggest other explanations for the discrepancies?**

**R Script**:  [WhiteHouse-WAVES.R](https://web.archive.org/web/20141222165725/http:/watchdoglabs.org/data/WhiteHouse/2012-11-30/WAVES/WhiteHouse-WAVES.R" \t "_blank" \o "WhiteHouse-WAVES.R)

***Series about analyzing White House visitor data***

* [Background information on White House visitor data](https://web.archive.org/web/20141222165725/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, WatchdogLabs.org, Dec. 11, 2012.**
* [White House Worker and Visitor Entry System Data Fields](https://web.archive.org/web/20141222165725/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](https://web.archive.org/web/20141222165725/http:/watchdoglabs.org/blog/2012/12/14/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/20141222165725/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/20141222165725/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/20141222165725/http:/watchdoglabs.org/blog/2013/08/02/latest-white-house-visitor-records-show-13-more-shulman-visits-in-2009/), WatchdogLabs.org, Aug. 2, 2013.