Earl F Glynn  
UMKC Center for Health Insights

2015-01-05

# Background

**National Plan and Provider Enumeration System** (NPPES)

**National Provider Identifier** Standard (NPI)

<http://www.cms.gov/Regulations-and-Guidance/HIPAA-Administrative-Simplification/NationalProvIdentStand/DataDissemination.html>

# File Download

NPI Downloadable File: **Full Replacement Monthly NPI File**

<http://nppes.viva-it.com/NPI_Files.html>

**November 2014 file**: 4,441,495 records each with 329 variables

<http://nppes.viva-it.com/NPPES_Data_Dissemination_November_2014.zip>

**December 2014 file**: 4,456,577 records each with 329 variables

<http://nppes.viva-it.com/NPPES_Data_Dissemination_December_2014.zip>

# File Structure

Overview of raw data file structure:

|  |  |
| --- | --- |
| Subset | Variables |
| NPI key | 1 |
| Base info 1 | 46 |
| Taxonomy License (15 x 4) | 60 |
| Other Identification (50 x 4) | 200 |
| Base info 2 | 7 |
| Taxonomy Group (15 x 1) | 15 |
| *total* | 329 |

Taxonomy License and Taxonomy Group data were combined in December 2014.

# Relational Tables

The raw file was converted to relational tables.

**NPPES/NPI database tables**

**N**

Taxonomy License

NPI key

*up to 15 sets of …*

Healthcare Provider Taxonomy Code

Provider License Number

Provider License Number State Code

Healthcare Provider Primary Taxonomy Switch

Healthcare Provider Taxonomy Group

**1**

NPPES info

NPI key

53 variables

**1**

5,320,425 records

4,456,577 records

**N**

Other Identifier

NPI key

*up to 50 sets of …*

Other Provider Identifier

Other Provider Identifier Type Code

Other Provider Identifier State

Other Provider Identifier Issuer

4,912,480 records

**NPPES info table**

|  |
| --- |
| NPI key |
|  |
| Entity Type Code  Replacement NPI  Employer Identification Number (EIN)  Provider Organization Name (Legal Business Name)  Provider Last Name (Legal Name)  Provider First Name  Provider Middle Name  Provider Name Prefix Text  Provider Name Suffix Text  Provider Credential Text  Provider Other Organization Name  Provider Other Organization Name Type Code  Provider Other Last Name  Provider Other First Name  Provider Other Middle Name  Provider Other Name Prefix Text  Provider Other Name Suffix Text  Provider Other Credential Text  Provider Other Last Name Type Code  Provider First Line Business Mailing Address  Provider Second Line Business Mailing Address  Provider Business Mailing Address City Name  Provider Business Mailing Address State Name  Provider Business Mailing Address Postal Code  Provider Business Mailing Address Country Code (If outside U.S.)  Provider Business Mailing Address Telephone Number  Provider Business Mailing Address Fax Number  Provider First Line Business Practice Location Address  Provider Second Line Business Practice Location Address  Provider Business Practice Location Address City Name  Provider Business Practice Location Address State Name  Provider Business Practice Location Address Postal Code  Provider Business Practice Location Address Country Code (If outside U.S.)  Provider Business Practice Location Address Telephone Number  Provider Business Practice Location Address Fax Number  Provider Enumeration Date  Last Update Date  NPI Deactivation Reason Code  NPI Deactivation Date  NPI Reactivation Date  Provider Gender Code  Authorized Official Last Name  Authorized Official First Name  Authorized Official Middle Name  Authorized Official Title or Position  Authorized Official Telephone Number |
|  |
| Is Sole Proprietor  Is Organization Subpart  Parent Organization LBN  Parent Organization TIN  Authorized Official Name Prefix Text  Authorized Official Name Suffix Text  Authorized Official Credential Text |

Because of Access database size issues, fields in blue above are loaded into a separate Access table.

The following summarizes the repeated groups from the original raw data file.

**Taxonomy License Sets** (up to 15 sets)

5,320,425 records

Counts by number of sets

**1 2 3 4 5 6 7 8 9 10**

4383720 616710 173439 57432 28708 17906 11842 8257 6016 4550

**11 12 13 14 15**

3535 2847 2250 1836 1377

Only 1377 providers specified data for all 15 possible taxonomy-license sets.

**Other Identifier** (up to 50 sets)

4,912,480 records

Counts by number of sets

**1 2 3 4 5 6 7 8 9 10**

1776975 1153615 752257 403529 250812 156081 105916 76037 56142 42105

**11 12 13 14 15 16 17 18 19 20**

31820 24074 18604 14467 11387 8958 7145 5651 4416 3224

**21 22 23 24 25 26 27 28 29 30**

1604 1204 966 784 621 497 417 358 316 285

**31 32 33 34 35 36 37 38 39 40**

248 222 201 179 160 146 135 125 111 102

**41 42 43 44 45 46 47 48 49 50**

92 83 75 66 63 56 49 40 34 26

Only 26 providers gave info for all 50 possible sets.

# Appendix. Processing Raw Data

## File Download

Script: **0-CMS-National-Provider-Identifier-Download.R**Output: Directory: DATA/yyyy-mm-dd  
Files: NPPES\_Data\_Dissemination.zip, **npidata\_20050523-yyyymmdd.csv**

## Recoding

The original raw data is quite bloated with millions of empty fields represented by a pair of quotation marks (""). To get rid of nearly 3 billion unnecessary quotes, the file will be re-written using tab delimiters instead of commas, but only after a small number of existing tab characters are removed. For details, see file **1-Recoding-notes.docx**.

Script: **1-Recode.bash** run in Linux virtual machine  
Input: npidata\_20050523-yyyymmdd.csv  
Output: **npidata\_20050523-yyyymmdd-notabs.csv**

## Reduce File Size

Remove almost 3 billion quotation marks and change the field separator from comma to tab.

Script: **2-CMS-National-Provider-Identifier-Rewrite.R**  
Input: npidata\_20050523-yyyymmdd-notabs.csv  
Outputs: npidata\_20050523-yyyymmdd-notabs.**txt**  
npidata-taxonomy-license.txt, npidata-other-identifier.txt  
**MASTER-NPPES-info.txt**

## Eliminate Repeating Groups

The original raw file is not "tidy" since two sets of different repeating groups are present. Let's change each repeating group into a separate table with a variable number of records instead of always allocating space for a fixed number of repeating groups.

Script: **3-CMS-NPI-split.R**  
Inputs: npidata-taxonomy-license.txt, npidata-other-identifier.txt  
Outputs: **MASTER-taxonomy-license.txt, MASTER-other-identifier.txt**

Key file summary [2014-12-10]

|  |  |  |
| --- | --- | --- |
| **Filename** | **Records** | **Size** |
| MASTER-NPPES-info.txt | 4,456,577 | 0.98 GB |
| MASTER-taxonomy-license.txt | 5,320,425 | 209 MB |
| MASTER-other-identifier.txt | 4,912,480 | 155 MB |

Record counts exclude header record.

## Extract Missouri Provider Addresses

The national data is nice for research purposes, but for now the focus is only on the Missouri providers. Let's create mailing and provider business addresses to be used in geocoding.

Script: **4-Missouri-Connections.R**  
Inputs: MASTER-NPPES-info.txt  
Outputs: NPI-Missouri-Connections-All.txt, NPI-Missouri-Practices.txt,  
NPI-Missouri-Practices-Mailing-Addresses.txt, **NPI-Missouri-Practices-Practice-Addresses.txt**