```
================================================================================
README FILE TO ACCOMPANY FRAC FOCUS 1.0 and FRAC FOCUS 2.0 DATASETS
LEVELS L0, L1, L2, L3, and L4
Data permalink: http://dx.doi.org/10.7910/DVN/EFNV5J
Datasets to accompany Energy Policy article at: 10.1016/j.enpol.2015.11.002
Revision History:
        ***V3 12 December 2015
        -added prepublication version of Energy Policy paper located at
        10.1016/j.enpol.2015.11.002
        -minor updates to methods paper
        ***V2 2 December 2015
        -added L4 .txt files by state, including dividing Texas by Job Start Year
        -renamed files with Level preceding name for more logical organization
        -Minor updates to methods paper, including addition of data link address.
        ***V1 21 July 2015
Konschnik, K (kkonschnik@law.harvard.edu) and Dayalu, A (adayalu@seas.harvard.edu)
--------------------------------------------------------------------------------
Contents:
(1) Level 0 Data Description
(2) Level 1 Data Description
(3) Level 2 Data Description
(4) Level 3 Data Description
(5) Level 4 Data Description
--------------------------------------------------------------------------------
Notes:

i.      **********************************IMPORTANT NOTE*******************************
        We recommend use of L2 data for analysis of "AS-IS" FracFocus
        data. L0 and L1 datasets are provided for information only, and contain
        various formatting issues that prevent smooth reading and analysis. The L2
        datasets addressed those issues while preserving the original
        FracFocus SQL-database content.

        L3 and L4 products are derived from the L2 dataset and contain additional
        fields we found necessary to accomplish our specific analyses; these may or may
        not be useful to others. In addition, the L3 and L4 products are SUBSETS
        of the original FracFocus dataset (at the L2 level) based on
        our analysis QA criteria of a selecting the latest (unique) form for a
        particular job occurring between January 1, 2011 and April 13, 2015.
        ***************************************************************************

ii.     Original data available in SQL format from http://fracfocus.org/data-download

iii.    The data contained in these files is based on data received directly from the
        Groundwater Protection Council (GWPC) on April 14, 2015. It contains data
        from forms submitted through April 13, 2015.

iv.     For details on data processing see the accompanying Methods Paper, “FracFocus
        Chemical Disclosure Registry 1.0 and 2.0 Data Conversion and Cleaning Methods
        Paper” (Dayalu and Konschnik 2015) accompanying these data.
================================================================================

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
1) Level 0 "L0" Data Description
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

The L0 dataset contains raw, pipe ("|) delimited output from querying the
RegistryUpload, Purpose, and Ingredients Tables in the original SQL database. There
are 12 "|"-delimited text files contained in this folder. With the exception of the
RegistryUpload table which was queried and output to a single file, all other fields
from the Ingredients and Purpose tables were queried and output column-by-column to
individual text files.

        ..........................................................................
1.      File(s) output from querying RegistryUpload table: FFApril_regupload.txt
        Contains the following fields:
        ++++++++++++   ++++++++++++++++++++++++++++++++++++++++++++++++++++++++
        FIELD NAME     DESCRIPTION
        ++++++++++++   ++++++++++++++++++++++++++++++++++++++++++++++++++++++++
        pKey:          Machine generated primary key for RegistryUpload table joins
        JobStartDate:  Date Job was reported as started, "YYYY-MM-DD HH:MM:SS" format
        JobEndDate:    Date Job was reported complete, "YYYY-MM-DD HH:MM:SS" format
        APINumber:     14-digit unique well identification number
        StateNumber:   2-digit API-system state code
        CountyNumber:  3-digit API-system county code
        OperatorName:  Name of well operator
        WellName:      Name of well
        WellType:      Type of well (e.g. Gas, Oil)
```

```
        Latitude:       Latitude of well, decimal degrees
        Longitude:      Longitude of well, decimal degrees
        Projection:     Geodetic Reference system (one of NAD27,Nad27,
                        NAD83, Nad83, WGS84)
        TVD:            True Vertical Depth of well, an imaginary perpendicular line
                        from the surface to the bottom of the borehole (UNITS?)
        TotalBaseWaterVolume: Volume of water injected into well
        TotalBaseNonWaterVolume: Volume of non-water components injected into well
        StateName:      Name of State in which well is located
        CountyName:     Name of County in in which well is located
        SubmitTimeStamp: Date form was submitted, "YYYY-MM-DD HH:MM:SS" format.
                         Can refer to amended time of submission, if form was
                         re-submitted at any point.
        FFVersion:      FracFocus Version (1 or 2) associated with form submission
        FederalWell:    True/False field
        Amended:        True/False/"" field
        OriginalSubmitTimeStamp: Original Date form was submitted, preserved
                                 in the event of amendments.
                                 "YYYY-MM-DD HH:MM:SS" format.

        ......................................................................
2.      File(s) output from querying Purpose table:
        FFApril_purpose_tradename.txt,
        FFApril_purpose_supplier.txt,
        FFApril_purpose_purpose.txt
        Contains the following fields:

        +++++++++++++   ++++++++++++++++++++++++++++++++++++++++++++++++++++++
        FIELD NAME      DESCRIPTION
        +++++++++++++   ++++++++++++++++++++++++++++++++++++++++++++++++++++++
        pKey:           Machine generated primary key for Purpose table joins
        pKeyRegistryUpload:     Machine generated primary key for RegistryUpload table
                                joins
        TradeName:      Chemical trade name
        Supplier:       Chemical Supplier Name
        Purpose:        Chemical Purpose

        ......................................................................
3.      File(s) output from querying Ingredients table:
        FFApril_ingreds_casnumber.txt,
        FFApril_ingreds_ingredcomment.txt,
        FFApril_ingreds_ingredmsds.txt,
        FFApril_ingreds_ingredname.txt,
        FFApril_ingreds_massingred.txt,
        FFApril_ingreds_pcthfjob.txt,
        FFApril_ingreds_pcthighadd.txt,
        FFApril_ingreds_thirdpartycalc.txt,

        +++++++++++++   ++++++++++++++++++++++++++++++++++++++++++++++++++++++
        FIELD NAME      DESCRIPTION
        +++++++++++++   ++++++++++++++++++++++++++++++++++++++++++++++++++++++
        pKey:           Machine generated primary key for Ingredients table joins
        pKeyPurpose:    Machine generated primary key for Purpose table joins
        CASNumber:      Chemical Abstract Service Registration Number, for
                        unique Chemical identification purposes
        IngredientComment:      Comment field
        IngredientMSDS: True/False field for availability of MSDS
        IngredientName:  Chemical Synonym(s)
        MassIngredient: Mass of reported ingredient (UNITS?)
        PercentHFJob:   Maximum Ingredient Concetration in Hydrualic Fracturing
                        fluid (% by mass)
        PercentHighAdditive:    Maximum Ingredient Concentration in Additive
                                (% by mass)
        ThirdPartyCalc: True/False field (purpose?)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
2) Level 1 "L1" Data Description
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

There are 12 L1 data files, mirroring those described in the L0 section above.
The L1 data sets are identical in content to L0 with the exception of removal
of newline "\n" instances. This was necessary for removal of newline fields often
erroneously appended to non-end-of-line fields. See the Methods Paper for details.
The L1 data file naming scheme is identical to that of the L0 files, except they
are all preceded with "L1_CLEAN_"

NOTE: The formatting of the L1 dataset is such that all lines are wrapped into
a single line. It is not recommended for use in analysis; it is an intermediate
```

```
data product.

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
3) Level 2 "L2" Data Description
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

There are 4 L2 data files: three "|"-delimited text files and one .RData format file.
The L2 cleanup process continued the formatting procedure to finally
output 3 tables for RegistryUpload, Purpose, and Ingredients fields.
See methods paper for details.

L2 .txt Files: Description
++++++++++++++++++++++++++
L2_ChemRegApril_regupload_clean_pipe_delim.txt:
        Final technical formatting of RegistryUpload table fields.
        No merging required as entire table was queried at once.
L2_ChemRegApril_purpose_clean_pipe_delim.txt:
        Final technical formatting of Purpose table fields. Merged in Rv3.2.0 using the
        Purpose table pKey as the join.
L2_ChemRegApril_ingreds_clean_pipe_delim.txt:
        Final technical formatting of Ingredient table fields. Merged in Rv3.2.0 using the
        Ingredients table pKey as the join.

L2 .RData File: Description
++++++++++++++++++++++++++
L2_ff_clean.RData:      Contains two R objects: reg_upload, ff2_clean
                        ,,,,,,,,,,,,,,
                        (1) reg_upload
                        Contains all header fields for FracFocus 1.0 and 2.0,
                        read in from ChemRegApril_L2_regupload_clean_pipe_delim.txt
                        Naming convention of fields preserved as in L0 and L1 data sets.
                        ,,,,,,,,,,,,,,
                        (2) ff2_clean
                        Contains ALL FracFocus 2.0 forms with associated Purpose and
                        Ingredients data.
                        Read in from ChemRegApril_L2_purpose_clean_pipe_delim.txt and
                        ChemRegApril_L2_ingreds_clean_pipe_delim.txt.
                        Purpose data merged with relevant RegistryUpload data by joining
                        across the RegistryUpload pKey.
                        Ingredients data merged with joined RegistryUpload/Purpose data
                        using the Purpose pKey.
                        Naming convention of fields preserved as in L0 and L1 data sets,
                        with the exception of the pKeys: all pKey fields were renamed
                        prior to the merging process to reflect table of origin
                        (one of "pKeyRegistryUpload","pKeyPurpose",
                        or "pKeyIngredients"). There is no longer a field named
                        simply "pKey".


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
4) Level 3 "L3" Data Description
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

There is a single .RData file. It is a subset of the L2 data, based on our analysis QA
criteria of a selecting the latest (unique) form for a particular job occurring between
January 1, 2011 and April 13, 2015. It also contains several additional fields,
created during our data standardization/cleanup process.

L3_ff_clean.RData:      Contains two R objects: forms_latest, ff2_forms_latest
                        ,,,,,,,,,,,,,,
                        (1) forms_latest
                        Contains all header fields as in L2 dataset, and additionally
                        contains the following NEW or MODIFIED fields:
                        +++++++++++++++ ++++++++++++++++++++++++++++++++++++++++++++++++++
                        FIELD NAME      DESCRIPTION
                        +++++++++++++++ ++++++++++++++++++++++++++++++++++++++++++++++++++
                        CASNumber       MODIFIED from L2 CASNumber field to remove spaces.
                                        This prevents correctly entered CASNumbers from
                                        being flagged as invalid if embedded spaces were
                                        present.
                        search_id_form: "JobStartDate_APINumber"; interchangeable with
                                        pKeyRegistryUpload as a unique form identification
                                        field at this L3 stage;
                                        added to aid in the form sub-selection process.
                        LatitudeClean:  Latitude, decimal degrees, converted to NAD83
                                        from NAD27 instances. NAD83 and WGS84 instances
                                        remained the same as in the Latitude field.
                        LongitudeClean: Longitude, decimal degrees, converted as described
```

in LatitudeClean field.
ProjectionClean:Geodetic reference system, pointing to datum of LatitudeClean and LongitudeClean fields. One of NAD83 or WGS84.
StateNameFromCoords: State Name based on LatitudeClean and LongitudeClean fields, from querying the FCC database located at https://www.fcc.gov/developers/census-block-conversions-api
StateAbbFromCoords: State Abbreviation based on LatitudeClean and LongitudeClean fields, from querying the FCC database as above
CountyNameFromCoords: County Name, based on LatitudeClean and LongitudeClean fields, from querying the FCC database as above
CountyCodeFromCoords: County Code, based on LatitudeClean and LongitudeClean fields, from querying the FCC database as above
StateNameFromAPI: State Name, derived from the first two digits of the user-input API number. NOTE: this is a different number system from the Federal Information Processing Standard state code (FIPS) code. The different system was accounted for in the conversion process.

,,,,,,,,,,,,,,,
(2) ff2_forms_latest
FracFocus 2.0 header, Purpose, and Ingredient data. In addition to the fields described above it also contains a single NEW field:

FIELD NAME	DESCRIPTION
SupplierClean	Standardized chemical supplier naming scheme for the top suppliers -- Halliburton, Baker Hughes, Schlumberger, and Multi-Chem (Halliburton). All entries in this field that correspond to other Suppliers are designated as NA

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
5) Level 4 "L4" Data Description
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

L4 data is derived from the L3 dataset. It also contains several additional fields, created during our data analysis process.

There are two pipe "|" delimited text files of the (1) Header data and (2) FracFocus2.0 all data. There is also a single .RData file with the Header and FracFocus2.0 datasets. In addition, there are pipe "|" delimited text files written out containing header and ingredient data FOR EACH STATE that has submitted data to FracFocus in the time period covered by the original dataset.

L4_ff1and2_header.txt: FF1.0 and FF2.0 Header Data. Contents are as below, detailed in the .RData L4 file description.

L4_ff2_all.txt: FF2.0 all data. Contents are as below, detailed in the .RData L4 file description.

L4_ff_latest_forms.RData: Contains two R objects: forms_latest, ff2_forms_latest

L4_ST_Header_Data.txt: Contains state-specific header data for FF1.0 and FF2.0

L4_ST_FF2_Ingredients_Data.txt: Contains state-specific FF2.0 data where ST is the two letter State abbreviation

********** IMPORTANT NOTE FOR TEXAS DATA ONLY **********
|Due to the volume of its data, ONLY TEXAS datasets are broken |
|up by years, where the year refers to the reported year of the Job |
|Start Date. The years are 2011 through 2015. |
|eg., TX_2011_Header_Data_L4.txt; TX_2011_FF2_Ingredients_Data_L4.txt |
| |
|The maximum number of rows for Texas is |
|in the TX_2014_FF2_Ingredienbts_Data_L4 file (504,629 rows) |
|which should be read completely by Excel versions beginning in 2007. |
****

,,,,,,,,,,,,,,
(1) forms_latest
Contains all header fields as in L3 dataset, and additionally
contains the following NEW field:

FIELD NAME	DESCRIPTION
StateOK	True/False field based on data QA requirement of no state locational ambiguity. Flagged as TRUE if StateNameFromAPI matches the independent field StateNameFromCoords. Flagged FALSE if otherwise.

,,,,,,,,,,,,,,
(2) ff2_forms_latest
FracFocus 2.0 header, Purpose, and Ingredient data. In addition to
the field described above it also contains several NEW fields:

FIELD NAME	DESCRIPTION
CASNumberClean:	All leading zeroes removed from mathematically valid CASNumber fields. All other entries in the CASNumber field labeled as NA in CASNumberClean.
CASLabel:	Flagged relative to original CASNumber field. One of "Valid","Invalid","Confidential","Proprietary", "Trade Secret", "N/A", or "Systems Approach"
WithheldFlag:	True/False field. Flagged TRUE for CASLabel entries corresponding to "Confidential","Proprietary", "Trade Secret",or "N/A". Flagged FALSE for all others.
Systems_Approach:	True/False/NA field. Flagged TRUE if entry was designated as Systems Approach AND supplier was a non-NA entry in the SupplierClean field. Flagged FALSE if the ingredient record was not a systems approach type entry for the non-NA entries in SupplierClean. For all other suppliers (NA in the SupplierClean field), Systems_Approach was flagged as NA (i.e., untested).
IngredientNameClean:	Standardized ingredient name obtained by database searches of CASNumberClean entries.

========== END ==========