EPA/601/R-14/006 I March 2015 I www.epa.gov/hfstudy

United States
Environmental Protection
Agency

Analysis of Hydraulic Fracturing Fluid Data from the FracFocus Chemical Disclosure Registry 1.0:

Data Management and Quality Assessment Report

United States Environmental Protection Agency
Office of Research and Development

[This page intentionally left blank.]

Analysis of Hydraulic Fracturing Fluid Data from the FracFocus Chemical Disclosure Registry 1.0: Data Management and Quality Assessment Report

U.S. Environmental Protection Agency
Office of Research and Development
Washington, DC

March 2015
EPA/601/R-14/006

Disclaimer

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Preferred Citation: U.S. Environmental Protection Agency. 2015. Analysis of Hydraulic Fracturing Fluid Data from the FracFocus Chemical Disclosure Registry 1.0: Data Management and Quality Assessment Report. Office of Research and Development, Washington, DC. EPA/601/R-14/006.

Table of Contents

List of Tables

List of Figures

Preface

The U.S. Environmental Protection Agency (EPA) is conducting a *Study of the Potential Impacts of Hydraulic Fracturing for Oil and Gas on Drinking Water Resources*. The study is based upon an extensive review of the literature; results from EPA research projects; and technical input from state, industry, and non-governmental organizations, as well as the public and other stakeholders. A series of technical roundtables and in-depth technical workshops were held to help address specific research questions and to inform the work of the study.

In Fiscal Year 2010, Congress urged the EPA to examine the relationship between hydraulic fracturing and drinking water resources in the United States. The EPA's *Plan to Study the Potential Impacts of Hydraulic Fracturing on Drinking Water Resources* was reviewed by the agency's Science Advisory Board (SAB) and issued in 2011. The *Study of the Potential Impacts of Hydraulic Fracturing on Drinking Water Resources: Progress Report*, detailing the EPA's research approaches and next steps, was released in late 2012 and followed by a consultation with individual experts convened under the auspices of the SAB.

This report, *Evaluation of Hydraulic Fracturing Fluid Data from the FracFocus Chemical Disclosure Registry 1.0: Data Management and Quality Assessment Report*, is the product of one of the research projects conducted as part of the EPA's study. It has undergone independent, external peer review, which was conducted through the Eastern Research Group, Inc. All peer review comments were considered in the report's development. The report has also been reviewed in accordance with agency policy and approved for publication.

The EPA is writing a state-of-the-science assessment that integrates a broad review of existing literature, results from peer-reviewed EPA research products (including this report), and information gathered through stakeholder engagement efforts to answer the fundamental research questions posed for each stage of the hydraulic fracturing water cycle:

- Water Acquisition: What are the possible impacts of large volume water withdrawals from ground and surface waters on drinking water resources?
- Chemical Mixing: What are the possible impacts of surface spills on or near well pads of hydraulic fracturing fluids on drinking water resources?
- Well Injection: What are the possible impacts of the injection and fracturing process on drinking water resources?
- Flowback and Produced Water: What are the possible impacts of surface spills on or near well pads of flowback and produced water on drinking water resources?
- Wastewater Treatment and Waste Disposal: What are the possible impacts of inadequate treatment of hydraulic fracturing wastewaters on drinking water resources?

The state-of-the-science assessment is not a human health or an exposure assessment, nor is it designed to evaluate policy options or best management practices. As a Highly Influential Scientific Assessment, the draft assessment report will undergo public comment and a meaningful and timely peer review by the SAB to ensure all information is high quality.

Acknowledgements

The EPA would like to acknowledge the Ground Water Protection Council and the Interstate Oil and Gas Compact Commission for providing data and information for this report. Assistance was provided by The Cadmus Group, Inc., under contract EP-C-08-015. The contractor's role did not include establishing agency policy.

List of Acronyms

API	American Petroleum Institute
CASRN	Chemical Abstracts Service Registry Number
CBI	Confidential Business Information
CSV	Comma-Separated Values
EIA	U.S. Energy Information Administration
EPA	U.S. Environmental Protection Agency
FIPS	Federal Information Processing Standards
GIS	Geographic Information System
GWPC	Ground Water Protection Council
ID	Identification
IOGCC	Interstate Oil and Gas Compact Commission
NAD	North American Datum
PDF	Portable Document Format
QA	Quality Assurance
TVD	True Vertical Depth
USGS	U.S. Geologic Survey
WGS	World Geodetic System
XML	Extensible Markup Language

[This page intentionally left blank.]

1. Introduction

This report describes the procedures used to develop a database from data submitted to the FracFocus Chemical Disclosure Registry (subsequently referred to as "FracFocus") by well operators. The resulting project database was used to conduct the analyses described in the *Analysis of Hydraulic Fracturing Fluid Data from the FracFocus Chemical Disclosure Registry 1.0* (subsequently referred to as the "data analysis report;" US EPA, 2015).[1] This data management report can be used in conjunction with the project database and data analysis report to reproduce the results presented in the data analysis report and to conduct additional analyses, if desired.

2. Source Data

FracFocus is a publicly accessible website (www.fracfocus.org) managed by the Ground Water Protection Council (GWPC) and the Interstate Oil and Compact Commission (IOGCC) where oil and gas production well operators can disclose information about the composition of hydraulic fracturing fluids at individual wells.[2] Disclosures included in the project database were submitted to FracFocus by well operators using the FracFocus 1.0 format and were provided in portable document format (PDF) to the U.S. Environmental Protection Agency (EPA) by the GWPC in March 2013.[3] The PDF files were converted to Extensible Markup Language (XML) and parsed into a Microsoft Access database (Microsoft Corporation, 2012). Reviews of data quality were conducted on the project database to ensure that the results from analyses of the project database reflect the data contained in the original PDF disclosures, while identifying obviously invalid or incorrect data to exclude from analyses.

The source data provided by the GWPC were a bulk archive of 39,136 disclosures in PDF format that were submitted to the FracFocus 1.0 website prior to March 1, 2013. Each disclosure was initially submitted by the well operator to FracFocus in the form of a Microsoft Excel spreadsheet and contained information on one production well that was hydraulically fractured with a single fracture date. Each Excel spreadsheet was then converted into a PDF file by the FracFocus website.

3. Database Development

The initial development of the project database involved data conversion of disclosures from PDF format to XML files, parsing to extract information, and incorporation of the resulting data into a

[1] The project database and the data analysis report are available at http://www2.epa.gov/hfstudy/published-scientific-papers.

[2] Prior to February 28, 2011, six of the 20 states with data in the project database began requiring operators to disclose chemicals used in hydraulic fracturing fluids to FracFocus (Colorado, North Dakota, Oklahoma, Pennsylvania, Texas, and Utah). Three other states started requiring disclosure to either FracFocus or the state (Louisiana, Montana, and Ohio), and five states required or began requiring disclosure to the state (Arkansas, Michigan, New Mexico, West Virginia, and Wyoming). Alabama, Alaska, California, Kansas, Mississippi, and Virginia did not have reporting requirements during the period of time studied in the data analysis report. Between February 5, 2011, and April 13, 2012, Pennsylvania required reporting to the state. As of April 14, 2012, Pennsylvania required reporting to both the state and FracFocus.

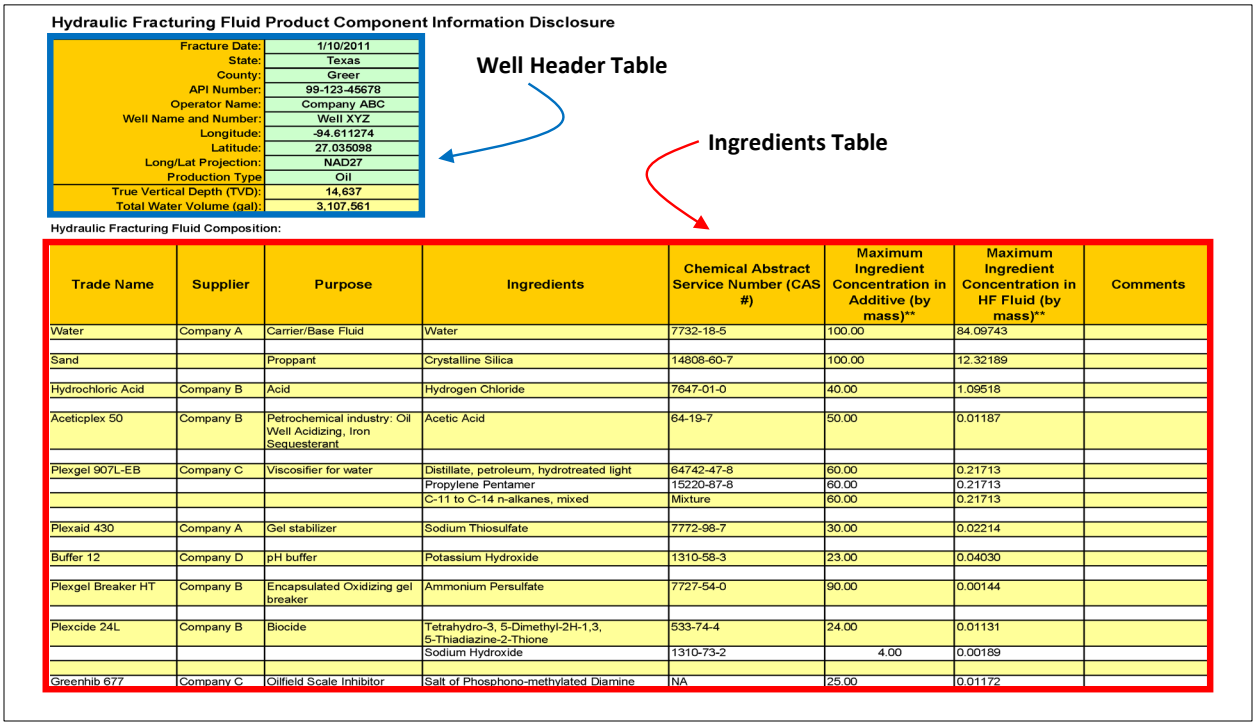
[3] FracFocus 2.0 became the exclusive disclosure mechanism in June 2013. More information on the FracFocus 1.0 FracFocus 2.0 formats may be found in the FracFocus 2.0 Operator Training materials available at http://fracfocus.org/node/331.

Microsoft Access database. The subsequent steps to conduct quality assurance (QA) and the resulting tables and fields that are suitable for data analysis are described in Sections 4, 5, 6, and 7. In describing the database development in this report, underline formatting denotes table names, **bold** formatting denotes field names, and *italic* formatting denotes data values.

3.1. Downloading and Conversion

The GWPC prepared a complete archive of all FracFocus 1.0 PDF disclosures (files) uploaded through February 28, 2013, and transferred the archive to the EPA. Adobe Acrobat Pro X (Adobe Systems Incorporated, 2011) was then used to convert all 39,136 PDF files in the archive to XML 2003 spreadsheet (Microsoft Excel 2003 XML) files. The conversion was performed because it is inherently difficult to extract data from PDF files, which are intended to provide consistent visual presentation across devices rather than structured representation of data for parsing and extraction. Tables of information in PDF files, in particular, can present a challenge for conversion. The source Microsoft Excel files, as uploaded by the operators, contained data in tables. However, in a PDF file, a table is essentially a series of lines and characters positioned on a page that, when assembled by PDF-reading software, appear as a table to the end user. To obtain tabular information from a PDF file, the PDF was converted to XML file format, which allows discrete data to be sorted into specific fields so that the data can be manipulated during analysis.

Each FracFocus 1.0 disclosure contains two tables of information. Figure 1 shows an example of an individual well disclosure available to the public as a PDF. At the top of each disclosure is the well header table (outlined in blue in Figure 1), which contains the fracture date, well identifiers [i.e.,



Fracture Date:	1/10/2011
State:	Texas
County:	Greer
API Number:	99-123-45678
Operator Name:	Company ABC
Well Name and Number:	Well XYZ
Longitude:	-94.611274
Latitude:	27.035098
Long/Lat Projection:	NAD27
Production Type	Oil
True Vertical Depth (TVD):	14,637
Total Water Volume (gal):	3,107,561

Trade Name	Supplier	Purpose	Ingredients	Chemical Abstract Service Number (CAS #)	Maximum Ingredient Concentration in Additive (by mass)**	Maximum Ingredient Concentration in HF Fluid (by mass)**	Comments
Water	Company A	Carrier/Base Fluid	Water	7732-18-5	100.00	84.09743	
Sand		Proppant	Crystalline Silica	14808-60-7	100.00	12.32189	
Hydrochloric Acid	Company B	Acid	Hydrogen Chloride	7647-01-0	40.00	1.09518	
Aceticplex 50	Company B	Petrochemical industry: Oil Well Acidizing, Iron Sequesterant	Acetic Acid	64-19-7	50.00	0.01187	
Plexgel 907L-EB	Company C	Viscosifier for water	Distillate, petroleum, hydrotreated light	64742-47-8	60.00	0.21713	
			Propylene Pentamer	15220-87-8	60.00	0.21713	
			C-11 to C-14 n-alkanes, mixed	Mixture	60.00	0.21713	
Plexaid 430	Company A	Gel stabilizer	Sodium Thiosulfate	7772-98-7	30.00	0.02214	
Buffer 12	Company D	pH buffer	Potassium Hydroxide	1310-58-3	23.00	0.04030	
Plexgel Breaker HT	Company B	Encapsulated Oxidizing gel breaker	Ammonium Persulfate	7727-54-0	90.00	0.00144	
Plexcide 24L	Company B	Biocide	Tetrahydro-3, 5-Dimethyl-2H-1,3, 5-Thiadiazine-2-Thione	533-74-4	24.00	0.01131	
			Sodium Hydroxide	1310-73-2	4.00	0.00189	
Greenhib 677	Company C	Oilfield Scale Inhibitor	Salt of Phosphono-methylated Diamine	NA	25.00	0.01172	

Figure 1. Example FracFocus 1.0 disclosure.

American Petroleum Institute (API) number and well name], locational data, production type, true vertical depth (TVD) of the well, and the total water volume used to hydraulically fracture the well.[4] The ingredients table (outlined in red in Figure 1) provides information on the trade names of the additives used in the hydraulic fracturing fluids, the supplier, and additive purpose. Each additive contains one or more ingredients, and the ingredients table includes the chemical name and Chemical Abstracts Service Registry Number (CASRN) for each ingredient, as well as the maximum concentrations as a percentage by mass in the additive and in the hydraulic fracturing fluid.

3.2. Extraction and Parsing

A script was used to read the XML files, parse the relevant data, and compile those data into a useable format. The parsing script was written in Python 2.7 (Python Software Foundation, 2012) and uses the Beautiful Soup 4 library (Richardson, 2013) to read the XML files.

The script first locates and extracts the well header information for a given file. Generally, the fracture date appears first in a PDF, followed by other parameters in order. The script locates the first cell in the file that is of cell type "DateTime."[5] The script then reads the columns below the date with the assumption that the other well header fields are ordered as anticipated from the disclosure template provided to well operators. In some cases, text wrapping in the original PDFs will split values into multiple rows, resulting in extra header cells. To address this, the position of the longitude field, which is always a negative number for locations within the United States, is used as a "landmark" to recalibrate the ordering of data fields.

The script parses information from the ingredients table by locating individual columns of information and then reading cells in that column until the bottom of the table is reached. The bottom of the table is either the last row with more than one cell or the last row in the sheet. Columns are located by searching for text patterns that indicate the presence of a column header. In developing the script, the text patterns were refined based on experience; some operators represent the same column of information differently. For the data fields **Purpose** and **Trade Name** in the ingredients table of the disclosure (Figure 1), operators generally enter a value once to indicate that an additive trade name or purpose applies to all ingredients that follow (e.g., additive "Plexgel 907L-EB" in ingredients table of Figure 1). Thus, a purpose and trade name are applied to ingredients until a new trade name and purpose are encountered. Blank values in the purpose and trade name columns are replaced with the previous value as the column is parsed.

The parsing approach is highly sensitive to formatting. If an operator departed from the FracFocus 1.0 template when originally creating a disclosure, the disclosure may have been skipped or information from the disclosure may have parsed incorrectly. Most of the disclosures were prepared in a consistent format that enabled relatively easy parsing of data. However, some disclosures were uploaded using templates modified by the operators, with columns added or

[4] More information on the field descriptions may be found in Section 7.1.1.

[5] Adobe Acrobat identified apparent dates and standardized them automatically. The standardization in this dataset was later reversed, because Acrobat occasionally "standardized" non-date values.

removed, fields left blank, or invalid data entered. The modified disclosures were problematic during parsing and QA.

Table 1. Summary of parsing success.

Well header parsed	Ingredient table parsed	Number of disclosures	Percentage of disclosures
Yes or No	Yes or No	39,136	100%
Yes	Yes or No	38,530	98.5%
Yes	Yes	37,017	94.6%
Yes	No	1,513	3.87%
No	No	606	1.55%

Note: "Yes" and "No" indicate whether portions of the disclosures (well header or ingredient table) were successfully parsed. "Yes or no" indicates that the disclosure counts include disclosures that were parsed and those that were not.

As shown in Table 1, the well header table was successfully parsed from 98.5% of disclosures (38,530 of 39,136), and both the well header and ingredient tables were successfully parsed from 94.6% of disclosures (37,017 of 39,136).

3.3. Output Data Structure

The script parsed the resulting data into two comma-separated value (CSV) files that form the foundation of the project database. One file contains the well operator, well identifiers, production, and locational data from the well header; the other file contains the additive, additive purpose, chemical, and chemical concentration data from the ingredients table. The two-table structure was considered appropriate because a one-to-many relationship exists between the well header values for an individual disclosure and the multiple values from the ingredients table that correspond to that disclosure. The two tables are linked in the project database by a constructed unique identification (ID) field. The ID field is necessary because the combinations of **API Well Number** and **Fracture Date** for 228 disclosures were found to be duplicated in the dataset and, thus, cannot serve as unique identifiers. Unique disclosures—defined by the combination of **API Well Number** and **Fracture Date**—were selected from duplicate disclosures by choosing the file with the most recent modification date. The modification date associated with each PDF is not information found on the publicly available disclosure that may be downloaded from FracFocus. If two or more records shared the same values for **API Well Number** and **Fracture Date**, then the PDF file with the most recent modification date was flagged as the authoritative disclosure.

To maximize the transparency of the QA effort, the final database contains two versions of the data extracted from the FracFocus 1.0 disclosures. The first version contains data as originally parsed without any formatting, spelling corrections, or standardization—these tables are denoted with the "Original" prefix in their names. The values in these tables were taken directly from the CSV files produced by the parsing script and are stored verbatim as text. The second version contains data

after formatting, corrections, and standardization were performed—these tables are denoted with the "Qa" prefix. The "Qa" tables also contain fields describing the adjustments made to each disclosure and whether the values met QA criteria. The two-version structure enabled straightforward review of all changes and streamlined tracing of disclosures back to the source data.

The primary tables in the project database are as follows:

- OriginalWell. Well header data with verbatim (unadjusted) values as parsed to input data.
- QaWell. Well header data with minor adjustments applied, including fixed typographical errors, removal of extraneous characters, and corrections of obvious transpositions (e.g., latitude and longitude swapped, state and county swapped). Columns accompanying each set of well header values, also referred to as QA flag fields, describe adjustments made to the OriginalWell data and whether the data met QA criteria as included in the QaWell table.
- OriginalIngredient. Ingredients data with verbatim (unadjusted) values as parsed to input data.
- QaIngredient. Ingredients data with minor adjustments applied, including corrected formatting of CASRNs and standardized suppliers. Similar to the table QaWell, the QaIngredient table includes QA flag fields that describe the adjustments made and whether the data met QA criteria for inclusion in analyses.

Additional tables in the database supporting the QA efforts and data analyses include the following:

- IngredientNameStandardization. Ingredient names were standardized using a list of chemical names paired with CASRNs compiled by the EPA. These standardized names are used in the QaIngredient table.
- PurposeStandardization. Additive purpose names were standardized and applied to the QaIngredient table to correct for spelling capitalization, spaces, and punctuation for most purpose entries. Synonyms for proppants and base fluids are also identified in this table.
- PurposeCategorization. Categorization of related additive purposes was applied to the standardized purposes for ease of summarizing the data during analyses. Information from this table was used for queries in which summary information was compiled regarding additive purposes.
- TradeNameStandardization. Standardized additive trade names were applied to values in the **TradeName** field to correct for spelling, capitalization, spaces, and punctuation and are used in the QaIngredient table.
- OperatorStandardization. Standardized operator names were applied to values in the Operator data to consolidate different representations of operator names and are used in the QaIngredient table.
- StateRegulation. This table lists effective dates for state laws that either mandate disclosure of hydraulic fracturing chemicals to FracFocus, allow FracFocus as an alternative to

reporting to state agencies, or require reporting to state agencies. (This information was obtained through separate research and is not information reported by operators to FracFocus.)

- Counties. This table provides a listing of all counties in the United States by state, name, and Federal Information Processing Standards (FIPS) code. This table also includes a separate identifier for the five case study counties included in the data analysis report.
- CBISynonym. A list was compiled of terms interpreted to indicate confidential business information (CBI) in the **Chemical Name** and **Cas** fields of ingredient records. This table was used for analyses of ingredient data reported as CBI or an associated term (such as 'proprietary,' 'trade secret,' etc.).
- Proppants. This table provides a listing of solid materials associated with proppant-related additive purposes and indicates whether these materials should be excluded from additive ingredient analyses conducted for the data analysis report.[6] The table is not associated with any changes or standardizations in the QaIngredient table, but was referenced in queries for chemicals.
- ResinCoating. This list contains ingredients associated with proppant-related additive purposes; these are ingredients that are not minerals, but rather chemicals associated with resin coatings on proppants. The list was referenced in queries for the proppants and additive ingredients analyses discussed in the data analysis report and is not associated with any changes or standardizations in the QaIngredient table.
- WaterSourceTerm. This list of terms is interpreted to indicate water sources reported by operators in the **TradeName** and **Comments** fields that are included in the QaIngredient table. These terms were used for the water source analysis described in the data analysis report.
- UnparsedPDFs. This table lists the PDFs that were unable to be parsed. It is incorporated for transparency and reference.
- WaterSynonyms. This list contains variations of operator entries (e.g., in the **TradeName, Comments,** or **ChemicalName** fields in QaIngredient) that indicate water but no other descriptors for the water source for base fluids. This list was used in querying for water sources. An ingredient record could match a term on this list only if it did not already match a term in WaterSourceTerm.

Section 7 describes the specific data fields found in these tables. Sections 4, 5, and 6, respectively, discuss the incorporation of geospatial data into the database, the QA procedures for well locational data, and the standardization of chemical names.

[6] Additive ingredients are defined as ingredients reported for additives that have purposes other than base fluid or proppant.

4. Assignment of Hydrocarbon Regions to Disclosures

Operators reported the production type (oil or gas) on FracFocus 1.0 disclosures, but not the specific producing formation. To offer basic geologic context for the locations of the disclosures, the hydrocarbon regions underlying each disclosure's latitude and longitude coordinates were added to the QaWell table after conversion of the coordinates to the North American Datum 83 (NAD83) in Esri ArcGIS v. 10.1 geographic information system (GIS; Esri, 2012).

National-scale spatial data describing the areal extent of hydrocarbon regions are limited—local and regional studies are more common. Five publicly available datasets with national coverage were chosen to be spatially joined to well locations. The National Oil and Gas Assessment province boundaries shapefile was obtained from the U.S. Geological Survey (USGS; USGS, 1995), and shapefiles for coalbed methane basins, tight gas basins, and shale gas plays and basins were obtained from the U.S. Energy Information Administration (EIA; US EIA, 2007, 2011a, b). These datasets were used for general reference purposes and with the understanding that the boundaries are approximate and that production may not be occurring from the co-located play. The following text boxes describe the content of these databases and provide links to metadata and file download locations.

USGS Oil and Gas Provinces

Field name	USGSProvinces
Description	This dataset includes 71 very large oil and gas provinces delineated as part of the USGS's 1995 National Oil and Gas Assessment (USGS, 1995). Although this layer has coarse spatial resolution, it has the advantage of covering the entire lower 48 states plus Alaska, which means that (nearly) every disclosure in the project database will be located within a province.
Metadata	http://certmapper.cr.usgs.gov/geoportal/catalog/search/resource/details.page?uuid=%7B50B96CAA-20BD-4875-B3B2-BB3E1E6B1CD9%7D
Download	http://certmapper.cr.usgs.gov/data/noga95/natl/spatial/shape/pr_natlg.zip

EIA Shale Basins

Field name	ShaleBasin
Description	This dataset includes 32 major sedimentary basins that contain hydrocarbon-bearing shales and correspond to the translucent pink "Basins" in the EIA "Lower 48 States Shale Plays" map.
Metadata	http://www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/maps/maps.htm
Download	http://www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/maps/shalegasbasin.zip

EIA Shale Plays

Field name	ShalePlay
Description	This dataset includes 45 shale plays that correspond to the translucent orange "Current Plays" and yellow "Prospective Plays" in the EIA "Lower 48 States Shale Plays" map.
Metadata	http://www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/maps/maps.htm
Download	http://www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/maps/shalegasplay.zip

EIA Tight Gas Basins

Field name	TightGas
Description	This dataset includes 13 sedimentary basins that contain tight gas formations and correspond to the translucent pink "Basins" in the "Major Tight Gas Plays, Lower 48 States" map.
Metadata	http://www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/maps/maps.htm
Download	http://www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/maps/tightgasbasinplay.zip
File in ZIP archive:	TightGasBasins_EIA_June2010.shp

EIA Coalbed Methane Basins

Field name	CoalBed
Description	This dataset includes 98 sedimentary basins that contain coalbed methane and correspond to the translucent pink "Coal Basins, Regions & Fields" in the "Coalbed Methane Fields, Lower 48 States" map.
Metadata	http://www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/maps/maps.htm
Download	http://www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/maps/cbm_4shps.zip
File in ZIP archive:	CBMbasins_Reserv06_Prod06.shp

ArcGIS 10.1 software was used for the spatial join process. The ArcGIS for Desktop Basic license includes the "Spatial Join" geoprocessing tool, which is routinely used to link the attributes of multiple sets of spatial data. In this case, the hydrocarbon regions were "join features," and the disclosure locations were the "target features." The disclosure locations were determined by the latitude and longitude coordinates in the project database (after QA and conversion to NAD83

datum, as described in Section 5), corresponding to the **NAD83_Lon** and **NAD83_Lat** fields in the QaWell table of the database. The "Join Operation" parameter was "JOIN_ONE_TO_ONE" and the "Match Option" parameter was set to "INTERSECT," such that a disclosure must spatially intersect the join feature in order to be assigned its value.

The assignment of a hydrocarbon region to a disclosure record in the database is meant to give context to the disclosure location and is likely to be more reliable at the basin scale than at the play scale. Interpretations of the analysis results do not assume that the wells at the disclosure locations are producing from any of the co-located shale plays assigned by this spatial join. Another limitation in accurately assigning plays to the disclosure locations is that the EIA geospatial data do not include boundaries for tight sand plays or coalbed plays; only basin boundaries are available from EIA for these two types of unconventional plays. Therefore, in areas with stacked plays that include sands or coalbeds in addition to shales, it is not possible to determine whether the producing formation is a shale play or another formation based solely on the locational data and the spatial join. Also, comparable EIA geospatial data were not available for oil basins.

For 4,644 disclosures (12% of 38,530 disclosures), the disclosure locations were within the surface boundaries of two EIA shale plays (i.e., plays with active production that are at different depths in the same general surface area, also known as "stacked plays"). Because operators do not report the play or formation that is being hydraulically fractured, there is ambiguity regarding the appropriate formation for the disclosure. Although operators provided TVDs, it is unknown if some of these values may include lateral lengths. Given the limitations of the TVD data, they were not used to interpret formation in regions with stacked plays in cases of shale play overlap for a location. Therefore, the value assigned to the **ShalePlay** field of the QaWell table is a combination of the individual shale play names, delimited by forward slashes (e.g., Avalon-Bone Spring/Barnett-Woodford.

Arthur et al. (2014) and Carter et al. (2013) summarized data from FracFocus by plays by assuming that the geographic placement of disclosures approximated the geologic placement in popular production plays. Before using the same strategy to categorize results in the data analysis report, the accuracy of geospatial information in identifying plays associated with disclosures was assessed. The results of the spatial join were compared with analogous information from the commercial database DrillingInfo (DrillingInfo, 2011). Because the EIA geospatial data used for the spatial join included play-level boundaries for shales but not for tight sands or coalbeds (these were only delineated at the basin level), the comparison was limited to shales. DrillingInfo is populated using state databases and includes information on producing formations. It includes API well numbers that correspond to 7,761 disclosures in the project database. Of the 7,761 disclosures, 7,153 are co-located with the EIA boundaries for shale plays. Among these 7,153 disclosures, 83% had EIA shale play designations generally consistent with the operator-identified formations in DrillingInfo. Among the 17% of disclosures for which the EIA shale plays did not match the DrillingInfo formations, the mismatches generally occurred where there are stacked plays that include shales in addition to tight sands or coalbeds, and the producing formation is a sandstone, limestone, or coal-bearing formation.

At this time, the basin designations provide useful context for the project database, but shale play designations should be regarded with care in areas with stacked producing plays. Ultimately, the data were not summarized by play in the data analysis report to be consistent with the analysis of the data "as is."

5. Quality Assurance Process for Locational Data

The well header table in each disclosure includes three sources of locational data:

- State name and county name information, as stored in the **StateFFQA** and **CountyFFQA** fields, respectively, of the QaWell table.
- State and county information encoded in the first five digits of the API Well Number, as stored in the **APIFFQA** field of the QaWell table.
- Latitude and longitude coordinates in the well header, as stored in the **LatitudeFFQA** and **LongitudeFFQA** fields, respectively, of the QaWell table. The datum of the coordinates is stored in the **ProjectionFFQA** field of the QaWell table.

Because the three locational sources were easily available and comparable, the location was determined to have met QA criteria if all three locational data fields agreed.[7]

To validate the location of each disclosure, the state and county entries for each of these three fields were compared. First, the leading five digits from **APIFFQA** were converted to state and county names using lookup tables from the Society of Petrophysicists and Well Log Analysts (2010). Second, the states and counties that intersect the coordinates reported in the **LatitudeFFQA** and **LongitudeFFQA** fields were determined using ESRI ArcGIS 10.1 software. Due to the varying datums entered in the **ProjectionFFQA** field, four separate shapefiles were created:

- Disclosures with a *NAD83* projection were read into a point shapefile with the North American Datum of 1983 geographic coordinate system.
- Disclosures with a *WGS84* projection were read into a point shapefile with the World Geodetic System Datum of 1984 geographic coordinate system, and then transformed to NAD83 via the "NAD_1983_To_WGS_1984_1" datum transformation with the Project geoprocessing tool.
- Disclosures with a *NAD27* projection in the lower 48 United States were read into a point shapefile with the North American Datum of 1927 geographic coordinate system, and then transformed to NAD83 via the "NAD_1927_To_NAD_1983_NADCON" datum transformation with the Project geoprocessing tool.
- Disclosures with a *NAD27* projection with a **StateFFQA** listed as *Alaska* were read into a point shapefile with the North American Datum of 1927 geographic coordinate system, and

[7] Well locations in Alaska were not subject to county-level locational QA criteria, because the five-digit API well numbers in Alaska are not organized by counties. The coordinates for all disclosures from Alaska fall within the boundaries of the North Slope borough.

then transformed to NAD83 via the "NAD_1927_To_NAD_1983_Alaska" datum transformation with the Project geoprocessing tool.

Following datum transformations to NAD83, these four shapefiles were merged into a single shapefile using the Merge geoprocessing tool. The final latitude and longitude coordinates (after transformation to NAD83, if needed) were stored in the **NAD83_Lat** and **NAD83_Lon** fields, respectively, in the QaWell table.

To join state and county names to each disclosure location, the Spatial Join geoprocessing tool was used with the 2010 TIGER/Line shapefile of counties from the US Census Bureau (USCB, 2011) with the "Join Operation" parameter set to "JOIN_ONE_TO_ONE" and the "Match Option" parameter set to "INTERSECT." The resulting attribute table was exported to Microsoft Excel (Microsoft Corporation, 2002).

In Excel, the three sets of state and county locations were compared, resulting in six QA measures for the locational data. These comparisons were case-insensitive to avoid situations where, for example, the data values *Mckee* and *McKee* would not match. These comparisons also ignored spaces and hyphens to avoid situations where, for example, *Mc Kee* and *McKee* would not match. For each of the six comparisons, a QA flag field was added to the data table with *True* or *False* Boolean values:

- **StateMatchAPI_FF** indicates whether or not the API code for the state (**APIState**) matches the state reported in the well header table (**StateFFQA**).
- **StateMatchGIS_FF** indicates whether or not the state that contains the GIS-mapped disclosure location (**GISState**) matches the state reported in the well header table (**StateFFQA**).
- **StateMatchAPI_GIS** indicates whether or not the API code for the state (**APIState**) matches the state that contains the GIS-mapped disclosure location (**GISState).**
- **CountyMatchAPI_FF** indicates whether or not the API code for the county (**APICounty**) matches the county reported in the well header table (**CountyFFQA**).
- **CountyMatchGIS_FF** indicates whether or not the county that contains the GIS-mapped disclosure location (**GISCounty**) matches the county reported in the well header table (**CountyFFQA**).
- **CountyMatchAPI_GIS** indicates whether or not the API code for the county (**APICounty**) matches the county that contains the GIS-mapped disclosure location (**GISCounty**)**.**

Based on these six fields, two additional flags were added:

- **AllStateOK** is *True* if all three state comparison fields are *True*.
- **AllCountyOK** is *True* if all six state and county comparison fields are *True*.

Locational data were used in the data analysis report for analyses in which information was needed at the state or county level. The QA-related fields were used as appropriate to either exclude

disclosures that did not meet QA criteria from analyses or to categorize results with uncertain locational information.

6. Chemical Name Standardization

Ingredient names and CASRNs are entered by operators in the ingredients table, and the names can include a wide range of variations for a given ingredient, including synonyms, misspellings, different punctuations and formatting, and different alpha-numeric spacing. To identify ingredients used in hydraulic fracturing fluids, entries of both ingredient names and CASRNs were verified and standardized. The CASRNs were determined valid for analyses after being verified with the Chemical Abstracts Service (2014); ingredient records with invalid CASRNs were excluded from certain analyses presented in the data analysis report. Note that this approach assumes that the CASRN entered into the project database is correct.

Ingredient names for verified CASRNs were standardized using a list of unique chemical names paired with CASRNs developed by the EPA. This standardization was needed because of the above-noted range of presentations of ingredient names. Because the ingredient names were standardized, the names found in the data analysis report and the project database may differ from the names reported by operators in the original PDF disclosures.

The EPA used standardized chemical names from Appendix A in the agency's *Study of the Potential Impacts of Hydraulic Fracturing on Drinking Water Resources: Progress Report* (2012) for the EPA-standardized chemical names used in the project database and in this report.[8] Chemical name and structure quality control methods were used to standardize chemical names for CASRNs found in the project database, but not included in Appendix A of the *Progress Report*.[9] The same methods were used in the development of Appendix A of the *Progress Report* and ensure correct chemical names and CASRNs.

[8] Table A-1 in the *Progress Report*.

[9] In the majority of cases, valid CASRNs and the associated ingredient names in the project database were paired correctly for a given CASRN. If an ingredient name (whether specific or non-specific) did not match the CASRN reported by the operator, the CASRN was added to a chemical name standardization list and assigned a correct chemical name. The chemical standardization list consists of CASRNs paired with appropriate chemical names and was used to standardize chemical names in the project database based on the CASRNs reported by the operators. This process was undertaken because numerous synonyms and misspellings for a given chemical were present in the original data. Standardized, specific chemical names were identified using the EPA's Distributed Structure-Searchable Database Network (US EPA, 2013), the EPA's Substance Registry Services database (US EPA, 2014a), and the U.S. National Library of Medicine ChemID database (US NLM, 2014). Additional information on chemical name and structure quality control methods can be found at http://www.epa.gov/ncct/dsstox/ChemicalInfQAProcedures.html.

7. Data Field Descriptions

The sections below provide a listing and descriptions of the data fields in the project database tables.

7.1. Data Fields in Main Tables

The primary tables that contain the data from the disclosures are:

- OriginalWell
- QaWell
- OriginalIngredient
- QaIngredient

The two "Original" tables contain the data as parsed from the original PDF disclosures. In the two "Qa" tables, data have undergone basic standardization, and a series of QA flag fields has been established to facilitate analyses. Fields with "QA" or "flag" in their names are in the "Qa" tables.

7.1.1. Well Header Field Descriptions

This section lists the fields in the OriginalWell and QaWell tables, which contain information derived from the 38,530 disclosures with successfully parsed well headers. For convenience, these are grouped into relevant categories based on the well header source field.

Well ID

WellId	A unique identifier assigned to each disclosure that was parsed into the project database

Fracture Job Date

DateFF	The verbatim fracture date from the parsed disclosure	
DateFFQA	**DateFF** after minor editing to correct obvious typos, incorrect formatting, and remove invalid values	
DateFFflag	*OK*	38,277 disclosures (99.34%) with **DateFF** unchanged
	OK, formatted	2 disclosures (0.0052%) with **DateFF** reformatted to fix an obvious typo
	Early	222 disclosures (0.58%) with **DateFF** before *1/1/2011* (the first day of the study period), which resulted in a blank for these disclosures in the **DateFFQA** field
	Late	28 disclosures (0.073%) with **DateFF** after *2/28/2013* (the last day of the study period), which resulted in a blank for these disclosures in the **DateFFQA** field
	Unclear	1 disclosure (0.0026%) with **DateFF** that could not be read, which resulted in a blank for these disclosures in the **DateFFQA** field

State

StateFF	The verbatim state name from the parsed disclosure	
StateFFQA	**StateFF** after minor editing to correct obvious typos and differences in formatting	
StateFFflag	*OK*	33,699 disclosures (87.46%) with **StateFF** unchanged
	OK, misspelled	38 disclosures (0.099%) with **StateFF** corrected to fix an obvious typo
	OK, postal to full	4,793 disclosures (12.44%) with **StateFF** corrected to substitute postal code (e.g., *TX* changed to *Texas*)

County

CountyFF	The verbatim county name from the parsed disclosure	
CountyFFQA	**CountyFF** after minor editing to correct misspelled County names, remove extraneous "County" and "Parish" suffixes, and remove invalid values	
CountyFFflag	*OK*	36,758 disclosures (95.40%) with **CountyFF** unchanged
	OK, misspelled	563 disclosures (1.46%) with **CountyFF** corrected to fix an obvious typo
	OK, shortened	1,206 disclosures (3.13%) with **CountyFF** corrected to remove extraneous suffixes (e.g. *County*, *Parish*, *Borough*)
	Unclear	3 disclosures (0.0078%) with **CountyFF** that was omitted or otherwise erroneous, which resulted in a blank for these disclosures in the **CountyFFQA** field

API Well Number

APIFF	The verbatim API well number from the parsed disclosure	
APIFFQA	**APIFF** after minor editing to include leading zeroes and add hyphens	
APIFFflag	*OK*	29,168 disclosures (75.70%) with **APIFF** unchanged
	OK, formatted	9,352 disclosures (24.27%) with **APIFF** reformatted to include leading zeroes and add hyphens
	Different than filename	10 disclosures (0.026%) with **APIFF** different than the API well number embedded in the PDF filename

Operator

OperatorFF	The verbatim well operator from the parsed disclosure	
OperatorFFQA	**OperatorFF** after minor editing to aggregate synonymous and misspelled operator names	
OperatorFFflag	*OK*	9,935 disclosures (25.79%) with **OperatorFF** unchanged
	OK, mapped	28,595 disclosures (74.21%) with **OperatorFF** changed to a synonym based on the OperatorStandardization table

Well Name

NameFF	The verbatim well name from the parsed disclosure	
NameFFQA	Matches **NameFF** because no values required editing	
NameFFflag	*OK*	38,530 disclosures (100.0%) with **NameFF** unchanged

Longitude

LongitudeFF	The verbatim longitude from the parsed disclosure	
LongitudeFFQA	**LongitudeFF** after minor editing to correct obvious typos and transpositions, and to remove invalid values	
LongitudeFFflag	*OK*	38,394 disclosures (99.65%) with **LongitudeFF** unchanged
	OK, lat/lon swapped	4 disclosures (0.010%) with **LongitudeFF** clearly transposed with latitude
	OK, nonnegative	129 disclosures (0.33%) with **LongitudeFF** erroneously non-negative but otherwise valid
	Unclear	3 disclosures (0.0078%) with **LongitudeFF** likely erroneous based on the resulting map location, which resulted in a blank for these disclosures in the **LongitudeFFQA** field

Latitude

LatitudeFF	The verbatim latitude from the parsed disclosure	
LatitudeFFQA	**LatitudeFF** after minor editing to correct obvious typos and transpositions, and to remove invalid values	
LatitudeFFflag	*OK*	38,518 disclosures (99.97%) with **LatitudeFF** unchanged
	OK, lat/lon swapped	4 disclosures (0.010%) with **LatitudeFF** clearly transposed with longitude
	OK, negative	5 disclosures (0.013%) with **LatitudeFF** erroneously negative but otherwise valid
	Unclear	3 disclosures (0.0078%) with **LatitudeFF** likely erroneous based on the resulting map location, which resulted in a blank for these disclosures in the **LatitudeFFQA** field

Projection

ProjectionFF	The verbatim projection (technically a datum) from the parsed disclosure	
ProjectionFFQA	Matches **ProjectionFF** because no values required editing	
ProjectionFFflag	*OK*	38,530 disclosures (100.0%) with **ProjectionFF** unchanged

Production Type (oil or gas)

TypeFF	The verbatim production type from the parsed disclosure	
TypeFFQA	Matches **Type FF** because no values required editing	
TypeFFflag	*OK*	38,530 disclosures (100.0%) with **TypeFFQA** unchanged

True Vertical Depth

DepthFF	The verbatim true vertical depth (in feet) from the parsed disclosure	
DepthFFQA	**DepthFF** after minor formatting to remove units, average ranges, and remove invalid values	
DepthFFflag	*OK*	37,721 disclosures (97.90%) with **DepthFF** unchanged
	OK, formatted	81 disclosures (0.21%) with **DepthFF** formatted to remove units and other extraneous characters
	Range	5 disclosures (0.013%) with **DepthFF** given as a range, which resulted in the **DepthFFQA** value being averaged from the minimum and maximum range values
	High	14 disclosures (0.036%) with **DepthFF** greater than 25,000 feet, the upper threshold identified by the EPA for reasonable depths, which results in a blank for these disclosures in the **DepthFFQA** field
	Low	5 disclosures (0.013%) with **DepthFF** less than 500 feet, the lower threshold identified by the EPA for reasonable depths, which results in a blank for these disclosures in the **DepthFFQA** field
	Not given	704 disclosures (1.83%) with **DepthFF** not reported, which results in a blank for these disclosures in the **DepthFFQA** field

Total Water Volume

VolumeFF	The verbatim total water volume (in gallons) from the parsed disclosure	
VolumeFFQA	**VolumeFF** after minor formatting to remove units and remove invalid values	
VolumeFFflag	*OK*	38,108 disclosures (98.90%) with **VolumeFF** unchanged
	OK, formatted	27 disclosures (0.070%) with **VolumeFF** formatted to remove units and other extraneous characters
	OK, revised	140 disclosures (0.36%) with **VolumeFF** revised due to altered header format
	Empty, revised	32 disclosures (0.083%) with **VolumeFF** removed due to altered header format
	High	11 disclosures (0.029%) with **VolumeFF** greater than 50 million gallons (upper threshold set by the EPA), which results in a blank for these disclosures in the **LatitudeFFQA** field
	Not given	133 disclosures (0.35%) with **VolumeFF** not reported, which results in a blank for these disclosures in the **LatitudeFFQA** field
	Unclear	79 disclosures (0.21%) with **VolumeFF** given but not valid numbers, which results in a blank for these disclosures in the **LatitudeFFQA** field

Duplication

APICount	In table QAWell, the number of disclosures with this API well number. A total of 2,283 disclosures (5.93%) shared an API well number with at least one other disclosure.
Authoritative	In table QAWell, *True* if the disclosure is the authoritative disclosure among a set of duplicates with the same **APIFFQA** and **DateFFQA**, as determined by the folder date or file creation date. A total of 38,301 disclosures (99.41%) matched are authoritative.

Locational Data from API Well Number

APIState	In table QAWell, the name of the State associated with the first two digits of the API Well Number in the **APIFFQA** field. The State associations were downloaded from http://www.spwla.org/technical/us-state-codes.
APICounty	In table QAWell, the name of the County associated with the first five digits of the API Well Number in the **APIFFQA** field. The County associations were downloaded from http://www.spwla.org/xls/counties.xls.

Locational Data from GIS Spatial Join of Longitude/Latitude Coordinates

NAD83_Lon	The **LongitudeFFQA** coordinate, after being converted to the NAD83 datum
NAD83_Lat	The **LatitudeFFQA** coordinate, after being converted to the NAD83 datum
GISState	The name of the state in which the **NAD83_Lat** and **NAD83_Lon** are located. Coordinates did not intersect a state in 56 disclosures (0.15%), resulting in blank values for **GISState** field.
GISCounty	The name of the county in which the **NAD83_Lat** and **NAD83_Lon** are located. Coordinates did not intersect a county in 56 disclosures (0.15%), resulting in blank values for **GISCounty** field.
USGSProvince	The name of the USGS Oil and Gas Province coincident with the disclosure's coordinates. Coordinates did not intersect a USGS province in 56 disclosures (0.15%), resulting in blank values for **USGSProvince** field.
ShaleBasin	The name of the EIA Shale Basin coincident with the disclosure's coordinates. Coordinates did not intersect a shale basin in 1,120 disclosures (2.91%), resulting in blank values for **ShaleBasin** field.
ShalePlay	The name of the EIA Shale Play coincident with the disclosure's coordinates. Coordinates did not intersect a shale play in 14,894 disclosures (38.66%), resulting in blank values for **ShalePlay** field.
TightGas	The name of the EIA Tight Gas Basin coincident with the disclosure's coordinates. Coordinates did not intersect a tight gas basin in 4,170 disclosures (10.82%), resulting in blank values for **TightGas** field.
CoalBed	The name of the EIA Coal Bed Methane Basin coincident with the disclosure's coordinates. Coordinates did not intersect a coalbed methane basin in 20,534 disclosures (53.29%), resulting in blank values for **CoalBed** field.

State Locational Matching

StateMatchAPI_FF	*True* if **APIState** matches **StateFFQA**. The two field values matched for 38,476 disclosures (99.86%).
StateMatchGIS_FF	*True* if **GISState** matches **StateFFQA**. The two field values matched for 38,390 disclosures (99.64%).
StateMatchAPI_GIS	*True* if **APIState** matches **GISState**. The two field values matched for 38,381 disclosures (99.61%).

County Locational Matching

CountyMatchAPI_FF	*True* if **APICounty** matches **CountyFFQA**. The two field values matched for 37,733 disclosures (97.93%).
CountyMatchGIS_FF	*True* if **GISCounty** matches **CountyFFQA**. The two field values matched for 36,894 disclosures (95.75%).
CountyMatchAPI_GIS	*True* if **APICounty** matches **GISCounty**. The two field values matched for 37,372 disclosures (96.99%).

Other locational fields

AllStateOK	*True* if all three **StateMatch** fields are true. The three field values matched for 38,359 disclosures (99.56%).
AllCountyOK	*True* if all three **StateMatch** and all three **CountyMatch** fields are true. The three field values matched for 36,754 disclosures (95.39%).

7.1.2. Ingredient Field Descriptions

This section lists the fields in the OriginalIngredient and QaIngredient tables, which provide information on additives and their ingredients, as well as base fluids and proppants.

IngredientId	The unique identifier added to each ingredient record that was parsed into the database
WellId	The unique identifier added to each disclosure that was parsed into the database
TradeName	The ingredient trade name. A number of trade name values are comma-joined lists of multiple trade names for the entire disclosure. Microsoft Access cannot store many of these long values in a text field, but converting to Memo would increase database size.
Supplier	The ingredient supplier. Supplier values (names) were standardized manually in QAIngredient.

Table continued on next page

Purpose	The purpose assigned to a particular ingredient. In table QAIngredient, purpose entries were standardized manually to correct for misspellings, punctuation, hyphenation, and capitalization.
ChemicalName	The original value parsed from the disclosures, in the OriginalIngredient table; or the standardized chemical name, where available, in the QaIngredient table
Cas	The CASRNs of the ingredient as parsed from the disclosures, in the OriginalIngredient table. In the QaIngredient table, CASRNs have been stripped of non-numeric characters and properly hyphenated, and CASRNs with invalid check digits have been removed.
EPAIngredientId	The identifier that links ingredient name standardization in the QAIngredient table with the IngredientNameStandardization table. Records for 796,692 ingredients were matched to an **EPAIngredientName**.
AdditiveConcentration	The original "maximum ingredient concentration in additive (% by mass)" parsed from FracFocus disclosures, in the OriginalIngredient table. In the QaIngredient table, entries expressed as a single decimal value were kept intact, while non-numeric values or ranges for 353,157 values were changed to Null.
FluidConcentration	The original "maximum ingredient concentration in hydraulic fracturing fluid (% by mass)," in the OriginalIngredient table. Entries expressed as a single decimal value were kept intact, while non-numeric values or ranges for 291,293 values were changed to Null.
Comments	Comments entered by the operator on the FracFocus disclosure. No changes were made to values in this field.
ValidTradeName	*True* if the trade name should be regarded as valid. This flag is set based on the TradeNameStandardization table. Values of **TradeName** appear to not be trade names for 252,361 ingredients; these have been flagged in the QAIngredients table as having an invalid trade name (value of *False*).
ValidPurpose	*True* if the purpose should be regarded as valid. This flag is set based on the PurposeStandardization table. Values of **Purpose** appear not be purposes for 204,123 ingredient records; these have been flagged in the QAIngredients table as having an invalid purpose (value of *False*).are clearly not purposes.
ValidAdditiveConcentration	*True* if **AdditiveConcentration** is between 0 and 100. For 356,789 ingredients, this field has been flagged in the QaIngredients table as *False* (invalid value).
ValidFluidConcentration	*True* if **FluidConcentration** is between 0 and 100. For 293,614 ingredients, this field has been flagged in the QaIngredients table as *False* (invalid value).
ValidCas	*True* if **Cas** matches a standardized ingredient in the IngredientNameStandardization table. For 433,753 ingredients, this field has been flagged in the QaIngredients table as *False* (invalid value).

7.2. Data Fields in Tables Associated with Standardizations

Several tables store the corrections and standardizations used to develop the QAWell and QAIngredient tables. These standardizations have been conservatively developed to facilitate data analysis.

7.2.1. Chemical Name Standardization

The following table lists the fields in the IngredientNameStandardization table. Ingredient names for verified CASRNs were standardized using a list of unique chemical names paired with CASRNs that was developed by the EPA (Section 6).

EPAIngredientId	The primary key for the table, which can be used to join the QaIngredient and IngredientNameStandardization tables
EPAIngredientName	The chemical name for the ingredient as determined by the EPA
Cas	The CASRN corresponding to an individual chemical. The EPA provided unique identifiers in the form of *NOCAS_XXXXX* (where *XXXXX* is a numerical identifier) for chemicals without CASRNs.

7.2.2. Operator Standardization Information

This section lists the fields of the OperatorStandardization table.

Original	The original operator name, found in the **Operator** field of the OriginalIngredient table. The **OperatorFF** field in OriginalWell was joined to this table using this field during the standardization process.
Standardized	The standardized name to use in the **Operator** field of the QaIngredient table

7.2.3. Trade Name Standardization

This section lists the fields of the TradeNameStandardization table, in which trade names were standardized to correct spelling and punctuation and evaluated to identify and flag entries that do not represent additives (e.g., numerical values, purposes, chemical names). Some fields were used in assigning a value to the **ValidTradeName** field in the QaIngredient table. Other fields provide additional categorization for reference.

ID	A unique identifier for each row in this table.
Multiple Entries in Trade Name Field	Checked if the trade name value appears to list multiple trade names. Some operators listed all additives used in one cell. This field is used to determine the value of the **ValidTradeName** field.
Ingredient (General name) - not proppant	Checked if the value appears to be an ingredient. This field is used to determine the value of the **ValidTradeName** field.
Purpose Name	Checked if the value appears to be an additive purpose. This field is used to determine the value of the **ValidTradeName** field.
Number that looks like possible concentration	Checked if the value appears to be a chemical concentration (possibly the result of parsing errors). This field is used to determine the value of the **ValidTradeName** field.
Possible CASRN	Checked if the value appears to be a CASRN. This field is used to determine the value of the **ValidTradeName** field.
Other	Checked if there appears to be another type of problem with the trade name value. This field is used to determine the value of the **ValidTradeName** field.
Count A, B, C, D, E or F	*1* if any of the above 6 fields are checked, otherwise *0*.
May or may not be Trade Name	Checked if it is not readily clear if the entry refers to something other than the trade name
Commodity	Checked if the value of the trade name is a commodity name (e.g., *water*)
Proppant (generic or trade name)	Checked if the value appears to indicate a proppant
Suggested spelling or punctuation correction	The standardized value of the **TradeName** field of the QaIngredient table
Trade Name as Listed in FracFocus	The original value of the **TradeName** field of the OriginalIngredient table. The **TradeName** field in OriginalIngredient was joined to this table using this field during the standardization process.

7.2.4. Ingredient Purpose Standardization

This section lists the fields of the PurposeStandardization table, in which purposes were evaluated to identify and flag entries that do not represent purposes (e.g., numerical values, chemical names, operator names). Some fields were used in assigning a value to the **ValidPurpose** field in the QaIngredient table. Other fields provide additional categorization for reference; the two fields

referring to proppants were used in querying for proppants and in excluding proppants from additive ingredient analyses.

ID	A unique identifier for each row in this table
Multiple Entries in Purposes Field	Checked if the additive purpose value appears to list multiple purposes. Some operators listed the purposes of all additives used in one cell. This field is used to determine the value of the **ValidPurpose** field.
Ingredient (General Name)(excludes HCl)	Checked if the value appears to be a chemical ingredient. This field is used to determine the value of the **ValidPurpose** field.
Commercial Product Name that doesn't include purpose and not IDd	Checked if the value appears to be a trade name of an additive. This field is used to determine the value of the **ValidPurpose** field.
Purpose Can Be Inferred from Product Name or From Another Entry	Checked if the purpose be inferred from an additive name or some other purpose entry for another ingredient record. This field is used to determine the value of the **ValidPurpose** field.
Item is Likely a Proppant	Checked if the value appears to indicate a proppant, even though it does not use a common identifying term such as *proppant* or list one of the chemical names *sand, silica,* or *quartz*. This field is used to determine the value of the **ValidPurpose** field.
Other	Checked if there is another type of problem with the additive purpose value. This field is used to determine the value of the **ValidPurpose** field.
Count B, C, D, E, F, or G	*1* if any of the above 6 fields are checked, otherwise *0*.
Proppant - uses word Proppant or other Identifying Term	Checked if the value appears to indicate a proppant, using the word *proppant* or listing one of the chemical names *sand, silica,* or *quartz* or other identifying term
Purpose corrected for caps, spacing, dashes, misspellings	The standardized value of the **Purpose** field of the QaIngredient table
Purpose as Listed in FracFocus	The original value of the **Purpose** field of the OriginalIngredient table. The **Purpose** field in OriginalIngredient was joined to this table using this field during the standardization process.
Related to Base Fluid	Checked if the additive purpose appears to be related to the base fluid
Related to Alternative Carrier	Checked if the additive purpose appears to be related to a non-water base fluid. The relationship was determined by observation and used for analysis of non-water base fluids.

7.3. Data Fields in Other Tables

Several additional tables have been added to the database with lists that were used to support the analyses described in the data analysis report.

7.3.1. Proppant Identification

This section contains information about the Proppants table, which lists solids (e.g., minerals, ceramics) associated with proppant-related purposes (as parsed from disclosures). Information in this table assisted with excluding the minerals used as proppants from analyses of additive ingredients.

ChemicalName	The chemical name of the proppant. The **ChemicalName** field in QaIngredient was joined to this table using this field to identify proppants.
Cas	The CASRN of the proppant
OK to exclude	Checked if the chemical can be excluded from the additive ingredient analyses

7.3.2. Resin Coating Identification

This section contains information about the ResinCoating table, which lists ingredients parsed from disclosures associated with the additive purpose of resin coatings. This list assisted in capturing the ingredients used for resin coatings on proppants in analyses of additive ingredients.

ChemicalName	The chemical name of the resin coating. The **ChemicalName** field in QaIngredient was joined to this table using this field to identify resin coatings.
Cas	The CASRN of the resin coating

7.3.3. CBI Identification

This section contains information about the CBISynonym table, which lists terms used to indicate that an operator has claimed CBI status for an ingredient in the **ChemicalName** and **Cas** fields of the OriginalIngredient table. This table was used for analyzing the numbers of ingredient records in the database that were listed by the operators as CBI.

Term	A term indicating CBI

7.3.4. Water Source Identification

This section contains information about the WaterSourceTerm table, which lists terms in the **TradeName** and **Comments** fields of the OriginalIngredient table that indicate the source of water used for the base fluid (e.g., fresh, recycled). This table was used to query the database for information on water sources.

Source	A term indicating a water source

7.3.5. Purpose Categorization

This section contains information about the PurposeCategorization table, which lists the categories of purposes as found in the **Purpose** field of the QaIngredient table. This table was used to group ingredients by purpose category.

Category	The category of the standardized purpose
Purpose	The standardized purpose

7.3.6. State Regulation Information

This section contains information about the StateRegulation table, which contains information about state reporting requirements. A single state may have multiple rows when regulations are amended.

ID	A unique identifier for each row in this table
State	The name of a state
Reporting Requirement Type	The recipient of required reporting, either the FracFocus registry (*FracFocus*), the state regulator (*State*), both FracFocus and the state (*FracFocus AND State*), or either FracFocus and the state (*FracFocus OR State*).
EffectiveDate	The effective date of the state regulation.
Effective Date within FF DB Timeframe?	Either *Y* if the date is between 1/1/2011 and 2/28/2013 or *N otherwise.*
Notes	Notes about the regulation, including relevant limitations.

7.3.7. County Information

This section contains information about the Counties table, which contains information about counties.

STATE	The state abbreviation
COUNTY	The full name of a county (e.g., *Clay County*)
FIPS	The county FIPS code
STATE_FIPS	The state FIPS code
CountyName	The short name of a county (e.g., *Clay*)
StateName	The name of a state
CaseStudy	Identifies whether the county is a focus county in the data analysis report

7.3.8. Water Synonyms

This section contains information about the WaterSynonyms table, which contains a list of synonyms for an unknown water source.

TradeName	A synonym for an unknown water source

7.3.9. Unparsed PDFs

This section contains information about the UnparsedPDFs table, which lists the 606 PDF files that could not be successfully parsed (Table 1).

PDFName	The PDF filename of the unparsed disclosure
API_Final	The API well number, as extracted from **PDFName**
Data Storage Error	Identifies 14 disclosures that GWPC indicated should be excluded from the project database because of a data storage error
State	The state in which the disclosure is located, based on the API well number

8. Summary

The project database was developed from PDF disclosures given to the EPA by the GWPC and submitted to the FracFocus Chemical Disclosure Registry 1.0 before March 1, 2013. Data from the PDF files were converted to XML format, parsed, and incorporated into a Microsoft Access database. The data in the project database were then subject to QA procedures to ensure that the results from analyses of the project database reflect the data contained in the original PDF disclosures, while identifying obviously invalid or incorrect data to exclude from analyses. A conservative approach was used in all data handling; no records were deleted and the original data remain in the project database. To improve the results of analyses, data have been subject to minimal standardization of operator names, trade names, and purposes, as well as standardization of chemical names according to CASRNs. The standardized entries are included in the two "Qa" tables. During QA work on the project database, data limitations were encountered, and QA flag fields were developed to identify agreement among locational data and instances of problematic data. During data analysis, database queries and subsequent calculations were structured to compensate for these limitations. Results of analyses conducted on the project database are presented in the *Analysis of Hydraulic Fracturing Fluid Data from the FracFocus Chemical Disclosure Registry 1.0* (US EPA, 2015).

References

Adobe Systems Incorporated. 2011. Adobe Acrobat X Pro 10.

Arthur, JD, Layne, MA, Hochheiser, HW, and Arthur, R. 2014. Spatial and Statistical Analysis of Hydraulic Fracturing Activities in US Shale Plays and the Effectiveness of the FracFocus Chemical Disclosure System. SPE Hydraulic Fracturing Technology Conference, The Woodlands, Texas, February 4-6. Society of Petroleum Engineers.

Carter, KE, Hakala, JA, and Hammack, RW. 2013. Hydraulic Fracturing and Organic Compounds - Uses, Disposal and Challenges. SPE Eastern Regional Meeting, Pittsburgh, Pennsylvania, August 20-22. Society of Petroleum Engineers.

Chemical Abstracts Service. 2014. Check Digit Verification of CAS Registry Numbers. Available at http://www.cas.org/content/chemical-substances/checkdig. Accessed April 21, 2014.

DrillingInfo, Inc. 2011. DI Desktop December 2011 download.

Esri,Inc. 2012. ArcGIS 10.1.

Microsoft Corporation. 2013. Excel 2013.

Microsoft Corporation. 2012. Access 2013.

Python Software Foundation. 2012. Python 2.7.

Richardson, L. 2013. Beautiful Soup 4.

Society of Petrophysicists and Well Log Analysts. 2010. API Standards Information. Available at http://www.spwla.org/technical/api-codes. Accessed April 21, 2014.

US Energy Information Administration (US EIA). 2007. Data for the Coalbed Methane Panels. Oil- and Gas-Related Maps, Geospatial Data, and Geospatial Software. Available at http://www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/maps/maps.htm. Accessed April 18, 2014.

US EIA. 2011a. Data for the Tight Gas Plays Map. Oil- and Gas-Related Maps, Geospatial Data, and Geospatial Software. Available at http://www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/maps/maps.htm. Accessed April 18, 2014.

US EIA. 2011b. Data for the US Shale Plays Map. Oil- and Gas-Related Maps, Geospatial Data, and Geospatial Software. Available at http://www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/maps/maps.htm. Accessed April 18, 2014.

US Environmental Protection Agency (US EPA). 2012. Study of the Potential Impacts of Hydraulic Fracturing on Drinking Water Resources: Progress Report. EPA 601/R-12/011. US Environmental Protection Agency, Washington, DC. 278 pages.

US EPA. 2013. Distributed Structure-Searchable Toxicity (DSSTox) Database Network. Available at http://www.epa.gov/ncct/dsstox/index.html. Accessed April 21, 2014.

US EPA. 2014. Substance Registry Services. Available at http://ofmpub.epa.gov/sor_internet/registry/substreg/home/overview/home.do. Accessed April 21, 2014.

US EPA. 2015. Analysis of Hydraulic Fracturing Fluid Data from the FracFocus Chemical Disclosure Registry 1.0. EPA/600/R-1/003. US Environmental Protection Agency, Washington, DC. 168 pages.

US National Library of Medicine (US NLM). 2014. ChemID Plus Advanced. Available at http://chem.sis.nlm.nih.gov/chemidplus. Accessed April 21, 2014.

US Census Bureau (USCB). 2011. Toplogically Integrated Geographic Encoding and Referencing (TIGER)/Line Shapefiles. Available at ftp://ftp2.census.gov/geo/tiger/TIGER2010/COUNTY/2010. Accessed September 16, 2013.

US Geological Survey (USGS). 1995. Province Boundaries shapefile. National Oil and Gas Assessment. Available at https://catalog.data.gov/dataset/1995-national-oil-and-gas-assessment-province-boundaries. Accessed April 18, 2014.

[This page intentionally left blank.]

EPA
United States
Environmental Protection
Agency