HARVARD LAW SCHOOL
Environmental Law Program
POLICY INITIATIVE

FracFocus Chemical Disclosure Registry 1.0 and 2.0

Data Conversion and Cleaning Methods Paper

Archana Dayalu and Kate Konschnik

1. Introduction

This paper documents the steps taken to develop a clean and standardized data set from data submitted to the FracFocus Chemical Disclosure Registry by oil and gas well operators. Researchers at Harvard Law School and Harvard University, School of Engineering and Applied Sciences (Harvard) used the resulting data set to conduct data quality and chemical disclosure analyses. This paper will enable others to reproduce our data set. In addition, we provide our data set for public use at http://dx.doi.org/10.7910/DVN/EFNV5J .

In March 2015, the United States Environmental Protection Agency (EPA) published data that the agency scraped from 38,530 FracFocus 1.0 disclosure forms. EPA published a "Data Management and Quality Assessment Report"[1] (EPA Report) in support of its "Analysis of Hydraulic Fracturing Fluid Data from the FracFocus Chemical Disclosure Registry 1.0"[2] (EPA Analysis), detailing a fairly extensive data correction and flagging endeavor. We did not correct and flag data in as many fields, because our subsequent analyses were more focused than EPA's. Where our field reviews overlapped with EPA's, we note where our method differs.

We adopted the notation format used in the EPA Report: we underline the names of the three data tables displayed on FracFocus forms,[3] **bold** field names within those tables, and *italicize* the data values entered into those fields.

2. Background

In 2010, the Groundwater Protection Council (GWPC) and the Interstate Oil and Gas Compact Commission (IOGCC) launched FracFocus, an online hydraulic fracturing chemical registry. The registry may be found at fracfocus.org. The registry provides public access to chemical information that well operators report using in a hydraulic fracturing job at a well.

Initially, information was submitted to FracFocus in a Portable Document Format (PDF) (FracFocus version 1.0, or FF 1.0). GWPC retained header data from these forms in Microsoft SQL but did not receive or store the substantive chemical information in this format. In November 2012, FracFocus introduced upgrades to the registry, including an option to submit data in eXtensible Markup Language (XML) (FracFocus 2.0, or FF 2.0). After June 1, 2013, XML became the exclusive data entry format. GWPC stores all FF 2.0 information in Microsoft SQL, and shares information with the public in PDF format.

---

[1] U.S. Environmental Protection Agency. 2015. Analysis of Hydraulic Fracturing Fluid Data from the FracFocus Disclosure Registry 1.0: Data Management and Quality Assessment Report. EPA/601/R-14/006. Washington, DC: U.S. Environmental Protection Agency.

[2] U.S. Environmental Protection Agency. 2015. Analysis of Hydraulic Fracturing Fluid Data from the FracFocus Chemical Disclosure Registry 1.0. EPA/601/R-14/003. Washington, DC: U.S. Environmental Protection Agency.

[3] Throughout this paper, we refer to each disclosure form as a "form." Note that the EPA Report and EPA Analysis refers to each disclosure form as a "disclosure." The terms can be read to mean the same thing.

On April 14, 2015, GWPC provided Harvard with a Microsoft SQL backup file containing all disclosures submitted on FF 1.0 and 2.0 forms and retained in SQL. The backup file contained 97,964 disclosure forms submitted through April 13, 2015.

As noted above, the SQL backup file only includes header table data for FF 1.0 forms.[4] The header data indicates basic information about the well and the fracturing job. For FF 2.0 forms, the SQL file includes the header table as well as two tables containing substantive information about the chemical ingredients used in the hydraulic fracturing fluid.

Section 3 details the contents of the SQL backup file and recounts our export of this data to text files. Section 4 describes our process for correcting formatting errors; for applying quality assurance (QA) criteria, to identify our analysis subset; and for standardizing state names, well coordinates, Chemical Abstract Service (CAS) numbers, and major chemical supplier names. Section 5 describes our final data set.

3. Step One: Exporting the FracFocus Data from the SQL backup database

As an initial step, we imported the SQL backup file of FF data into three text files, corresponding to the three tables contained in the backup file:

- RegistryUpload: This table contains the header fields located in the box at the top left-hand corner of a FracFocus form. This table includes data regarding the hydraulic fracturing job date, the location and unique (American Petroleum Institute, or API, Number) identifier for the well, the name of the well operator, the well's vertical depth, the type and volume of base fluid used to fracture the well, and the FracFocus version.
- Purpose: This table contains the **TradeName**, **Supplier**, and **Purpose** fields in the large table below the header table. These fields describe each additive in the hydraulic fracturing fluid. Each additive, in turn, may contain one or more ingredients.
- Ingredients: This table includes fields labeled **IngredientName**, **CASNumber** (a unique identifier of chemical compounds), **PercentHighAdditive** (the maximum ingredient concentration in the additive, as a percentage), and **PercentHFJob** (the maximum ingredient concentration across the entire hydraulic fracturing fluid, as a percentage).

As noted in Section 2, the SQL backup file contains RegistryUpload data for hydraulic fracturing jobs submitted on FF 1.0 and 2.0 forms but contains Purpose and Ingredient data only for jobs submitted on FF 2.0 forms.

We used SQL Server 2014 Evaluation Edition (CAB 32-bit downloaded from http://www.microsoft.com/en-us/evalcenter/evaluate-sql-server-2014) for the download process. Using the SQL Import/Export wizard, we queried the following and specified the "|" (pipe) as the column delimiter to minimize formatting inconsistencies between the delimiter and actual

[4] As noted in Section 1, EPA's data set contains chemical data for FF 1.0 forms, because EPA scraped this information from the PDF forms.

characters present in the fields. Similarly, we specified both "Line Feed" ({LF}) and "Carriage Return" ({CR}) for our row delimiters.

Each table has its own machine generated primary key (pKey) used for joining entries for common forms across tables. In addition, the Purpose and Ingredient tables have pKeys that enable "joins" with their respective parent tables (Purpose is joined to pKeyRegistryUpload and Ingredient is joined to pKeyPurpose). We queried the ChemRegApril.dbo.RegistryUpload table all at once because it had the fewest newline and other formatting issues. We queried the ChemRegApril.dbo.RegistryUploadPurpose and ChemRegApril.dbo.RegistryUploadIngredients tables column by column, with each query involving the relevant pKey variables. The code is reproduced below.

```
RegistryUpload TABLE QUERY
SELECT * FROM ChemRegApril.dbo.RegistryUpload

Purpose TABLE QUERY
SELECT pKey,pKeyRegistryUpload,TradeName FROM ChemRegApril.dbo.RegistryUploadPurpose
SELECT pKey,pKeyRegistryUpload,Supplier FROM ChemRegApril.dbo.RegistryUploadPurpose
SELECT pKey,pKeyRegistryUpload,Purpose FROM ChemRegApril.dbo.RegistryUploadPurpose

Ingredients TABLE QUERY
SELECT pKey,pKeyPurpose,IngredientName FROM ChemRegApril.dbo.RegistryUploadIngredients
SELECT pKey,pKeyPurpose,CASNumber FROM ChemRegApril.dbo.RegistryUploadIngredients
SELECT pKey,pKeyPurpose,PercentHighAdditive FROM ChemRegApril.dbo.RegistryUploadIngredients
SELECT pKey,pKeyPurpose,PercentHFJob FROM ChemRegApril.dbo.RegistryUploadIngredients
SELECT pKey,pKeyPurpose,IngredientComment FROM ChemRegApril.dbo.RegistryUploadIngredients
SELECT pKey,pKeyPurpose,IngredientMSDS FROM ChemRegApril.dbo.RegistryUploadIngredients
SELECT pKey,pKeyPurpose,ThirdPartyCalc FROM ChemRegApril.dbo.RegistryUploadIngredients
SELECT pKey,pKeyPurpose,MassIngredient FROM ChemRegApril.dbo.RegistryUploadIngredients
```

4. Step Two: Formatting and Standardizing the Data

We cleaned and standardized certain fields by programming in Bash (Linux) to produce Level 1 (L1) datasets and in Rv3.2.0 to produce Levels 2 through 4 (L2, L3, and L4) datasets. L1 and L2 cleanup involved removal of special characters introduced during the export process from SQL to flat text format. Apart from a single special character replacement described below, we made no modification to the original user-entered values of the data set at the L2 stage. Our aim was to maintain the integrity of the original data set and avoid subjective modifications. See Section 4.1 for details. L3 and L4 cleanup involved creation of new fields to contain standardized values of data in existing fields, where typographical errors might limit or confound analyses. In addition, L3 cleanup involved objective modification of the **CASNumber** field, necessary for ease of analysis. See Sections 4.2 and 4.3 for details.

4.1 L1 and L2 Cleanup: Basic formatting

When we exported SQL data or viewed the data through another program, we observed that newlines (identified by special character "\n", translated from our original SQL specification of {LF} as one criterion for delimiting rows) often erroneously appeared in non-end-of-line fields. This resulted in the addition of extraneous empty rows. To address this, in the L1 cleanup process we replaced the extraneous newline instances in the original pipe-delimited files exported from SQL with spaces (" "), using the Bash translate ("tr") command. We used these spaces in the L2 Cleanup process to identify newline instances and remove extraneous newlines.

We conducted L2 cleanup in Rv3.2.0. Specifically, after importing the L1-cleaned pipe-delimited text files into R, we:

- Removed a single extraneous empty row that was appended to the end of each imported table. Apparently, the Unix carriage return (^M, translated from {CR} in SQL) character present in the original text files at the end of each line (including the last line of the file), signaled to R to add a new line.
- Deleted the spaces appended to the start of each row from the L1 cleanup process. The **pKey** field represents the start of each row and does not ordinarily contain any spaces; removal of spaces therefore returned the field to its original configuration while preventing the addition of extraneous empty rows described above.
- Removed all non-ASCII characters (specifically, the Unicode space "\xa0").
- Removed combinations of carriage returns and tabs ("\r\t").

In addition to these universal formatting changes, we made formatting changes in the following specific fields during the L2 cleanup process:

- All L1 RegistryUpload table fields were imported into R as characters. This step was ensured that the *API Numbers* were not converted to scientific notation format.
- The **IngredientsName** field in the Ingredients table contained a single instance where a pipe ("|") was used to separate characters in an entry. R interpreted this as a field delimiter and wrapped the remainder of the field to a new line. The row was re-appended to its original line and separated with a "/" instead of the original "|" to prevent further confusion of field delimiters with actual field characters.
- The Ingredients and Purpose fields were re-joined using the appropriate primary key to master Ingredients and Purpose tables using the merge command in R.

Once the above steps were completed, the L2 cleaned RegistryUpload, Ingredients, and Purpose tables were written out to new pipe delimited text files.

4.2 L3 Cleanup: Standardization of Fields for Analysis

4.2.1 Defining Analysis Subset

At this stage, we identified the subset of FracFocus forms that we would use for our analyses. Here, we followed a similar approach to the one taken in the EPA Analysis. Specifically, we adopted the same two primary quality assurance (QA) criteria adopted by the EPA and reproduced from their report below; our modifications are in **bold** font.

> *Data that were **[…]** incorporated in the **[analysis]** must first pass two primary QA criteria to be included **[…]**: the combination of fracture date and API well number for each disclosure must be unique (i.e., no duplicates), and the fracture date must occur between January 1, 2011, and **April 13, 2015**. While duplicate disclosures from the same fracturing event (i.e., same API well number and same fracture date) were excluded from analyses, more than one disclosure for a given well was included if the fracture dates on the disclosures differed.*[5]

We created **submit_time_fracday** and **search_id_form** fields in the process of identifying those forms that passed the criteria. The **submit_time_fracday** field is a sequential number providing the fractional Julian days of the SubmitTimeStamp field since January 1, 1970. The **search_id_form** field is of the format "*JobStartDate_APINumber*". Following application of the criteria, our analysis subset contained 43,366 FF 1.0 forms and 53,083 FF 2.0 forms, for a total of 96,449 forms. Note that upon application of the two primary QA criteria, form categorization by the **search_id_form** field is identical to form categorization by **pKeyRegistryUpload** field.

4.2.2 Creation of Two Data Sets: Header Data Set, and FracFocus 2.0 Data Set

Next, we created two data sets from the analysis subset. The first data set contains all header data (all fields in the RegistryUpload table) for all forms in the subset, whether entered on FracFocus 1.0 or 2.0 forms.

For the second data set, we merged all FF 2.0 fields in the RegistryUpload, Ingredients, and Purpose tables from the forms in our analysis subset using the appropriate **pKey** fields. We anticipated using this data set for chemical disclosure analyses. Each table's **pKey** field was renamed to reflect the table of origin, a necessary step for the merge which requires common field names across tables in order to conduct a join operation. We standardized pKey field names where necessary to reflect the table of origin. Therefore, the FF 2.0 merged data set contains only three pKey fields, namely, **pKeyRegistryUpload**, **pKeyPurpose**, and **pKeyIngredients**. The FF 1.0 dataset contains only one relevant pKey field, **pKeyRegistryUpload**.

[5] EPA Analysis, *supra* n. 2, at 12.

4.2.3 State Name and Well Coordinate Standardization

A number of our proposed analyses involved comparisons of data quality and rates of disclosure between states. Therefore, we anticipated using the **StateName** field in the RegistryUpload table to categorize some of our results.

In FF 1.0, well operators manually entered data values into this field. The values are presented in various formats including misspellings, abbreviated versions of the state's full name, county names, and blanks. The variations complicate state-specific queries, and can make it difficult to determine where a well is located. To ease analysis, we created two standardized State Name fields, as described in the following paragraphs.

In FF 2.0 forms, the **StateName** field was automatically populated based on the first two digits of the *API well number* as entered by the operator.[6] This standardized the values in the **StateName** field, although it did not guarantee the resulting value accurately described the well's location. (If an *API well number* were incorrect, the wrong state code might appear and likewise prompt entry of the wrong value in the **StateName** field.)

In FF 1.0 and 2.0, well operators manually entered latitude and longitude coordinates to describe the location of the well. Moreover, they reported coordinates derived from three different geographic coordinate systems:

- North American Datum 1927 (NAD27), based on ground survey information gathered in the nineteenth century (this datum was used on 66.4% of all forms);
- North American Datum 1983 (NAD83), based on ground survey information gathered in the twentieth century and satellite information (used on 32.3% of all forms); and,
- World Geodetic System 1984 (WGS84), the most recently developed datum for use worldwide (used on 1.3% of all forms).

NAD83 coordinates are nearly identical to those derived from WGS84. However, coordinates reported under NAD27 could be hundreds of meters away from coordinates derived from NAD83 or WGS84.[7] Therefore, we converted all NAD27 values to

[6] The American Petroleum Institute (API) established a method for identifying all oil and gas wells in the United States. The 14-digit API number provides three basic pieces of information about a well – the first two digits indicate the state where the well is located; digits three to five indicate the county where it is located; a 5 digit number indicating the permit number associated with that well. Where there are multiple "kicks" or laterals off of that same well, there may be another 4 digits appended to the API number. The API number enables researchers to identify the number of unique wells located in the FracFocus database.

[7] U.S. Geological Survey. 2013. Teaching About and Using Coordinate Systems. Available: http://education.usgs.gov/lessons/coordinatesystems.pdf [accessed 3 July 2015].

NAD83. We created **LatitudeClean** and **LongitudeClean** fields to hold coordinates as reported in NAD83 and WGS84 or converted to NAD83.

We then created two standardized State Name fields:

a. **StateNameFromAPI**. Here we used the FF 2.0 approach to standardize all the FF 1.0 state names, automatically populating this field based on the first two digits of a well's API number.

b. **StateNameFromCoords**. We populated this field by entering the values from the **LatitudeClean** and **LongitudeClean** fields into the Federal Communications Commission (FCC) Census Block Conversions website[8] and identifying the resulting state name. (In addition, we entered the state abbreviation, county name, and county code into three new fields called **StateAbbFromCoords**, **CountyNameFromCoords**, and **CountyCodeFromCoords**, respectively).

We used the standardized State Name values to apply a secondary QA criterion to forms in our analysis subset: clarity as to the well's state location. Where a value in the **StateNameFromAPI** field did not agree with the value in the **StateNameFromCoords** field, we removed the form from the subset before each analysis. (These forms remain in the L4 subset available indexed by a logical **StateOK** field (see Sec 4.3.1). The data error analysis in our paper details this process.

4.2.4 CAS Number Standardization, Part I: Basic Formatting

At this stage, we removed all spaces from the **CASNumber** field prior to conducting any CAS number verification analyses.

4.2.5 Supplier Name Standardization

We anticipated comparing data reported by different chemical suppliers, to determine how often each company used the "systems approach" to report chemicals (a reporting format that disaggregates chemical ingredients from a particular chemical additive, thereby discouraging efforts to reverse engineer the products). Our systems approach analysis would focus on the three largest service companies and chemical suppliers: Baker Hughes, Halliburton, and Schlumberger.[9] Therefore, we only standardized **Supplier** field entries for these three companies and a Halliburton affiliate.

---

[8] https://www.fcc.gov/developers/census-block-conversions-api. An example query would be http://data.fcc.gov/api/block/find?format=json&latitude=28.35975&longitude=-81.421988&showall=true.

[9] On March 27, 2015, Halliburton stockholders adopted a merger agreement with Baker Hughes and approved the proposed combination of the two companies.

In FracFocus 1.0 and 2.0, well operators manually entered data values into the **Supplier** field manually. The values contain numerous typographical errors. The variations would complicate any query seeking to capture all of the wells associated with those companies.

We identified all of the variations of a company name, and changed each to the correctly spelled name of the company, with the first letter of each word capitalized. Table 1 lists all the variations that we encountered for each of these three companies in the original **Supplier** field along with the standardized name in a newly created **SupplierClean** field.

Table 1. **Supplier** Values for Top Four Suppliers

Supplier Field Values	Standardized **SupplierClean** Value
Hailliburton, Hailliburton , Haliburton, Halilburton, Hallburton, HALLBURTON, Halliberton, Hallibrton, Hallibruton, HALLIBURT, Halliburtion, halliburton, halliburton, Halliburton, Halliburton, Halliburton, Halliburton `, Halliburton, HALLIBURTON, Halliburton.com, Halliburton Energy Services, HALLIBURTON ENERGY SERVICES, Halliburton Energy Services, Inc., Hallibuton, Hallibutron, Hallinurton, Hallisburton, Hallivurton, Hallliburton, Halluburton, Hilliburton	*Halliburton*
Halliburton/Multi-chem, Halliburton - Multi-Chem, Halliburton / Multi-Chem, Halliburton Multi Chem, Halliburton, MultiChem, Halliburton/Multi Chem, Halliburton/Multi-Chem, HALLIBURTON/MULTI-CHEM, MulitChem, Mulit Chem, Mulit-Chem, Muliti-Chem, MultChem, Mult-Chem, Multi-Chame, multi-chem, Multichem, Multi-chem, MultiChem, Multi Chem, Multi Chem, Multi-Chem, Multi-Chem, MultiChem, MultiCHem, Multi-CHem, MULTICHEM, MULTI-CHEM, MULTI-CHEM, Multichem Chemicals, Multi-Chem Group LLC, Multi-Chem Group LLC, Multi=chem/Halliburton, Multi-chem/Halliburton, Multi-Chem (Halliburton), Multi-Chem/Halliburton, Multichems, Mulyi-Chem, Muti-Chem, Mutli-Chem	*Multi-Chem (Halliburton)*
Backer Hughes, Bake Petrolite, Baker, Baker Add-Frac, Baker Highes, Baker Highes, Baker Huches, Baker Huges, BAKER HUGES, Baker Hugher, baker Hughes, Baker hughes, BakerHughes, Baker Hughes, Baker Hughes, Baker Hughes, BAKER HUGHES, BAKER HUGHES, BAKER-HUGHES, Baker Hughes Company, Baker Hughes, Inc, Baker Hughes, Inc., Baker Hughs, Baker Hughtes, Baker Huhges, Baker Petrolite, BAKER PETROLITE, Bsker Hughes	*Baker Hughes*
Schlumberger, Schlumberger, SCHLUMBERGER, Schlumberger Technology Corporation, Schulmberger, Schumberger, Sclumberger, SCLUMBERGER	*Schlumberger*

4.3. L4 Cleanup: Standardization of Additional Fields During Analysis

4.3.1 Location Verification

We created a new logical field, **StateOK**, that was flagged as *TRUE* if the **StateNameFromCoords** fields matched the **StateNameFromAPI** field and *FALSE* if these two state name fields did not match.

4.3.2 CAS Number Standardization, Part II: Mathematical Verification

We created a new field, **CASLabel**, where non-numeric or incorrectly formatted CAS number entries were labeled as one of *Invalid*, *Confidential*, *Proprietary*, *Trade Secret*, *N/A*, or *Systems Approach*. See Table 2. We passed correctly formatted **CASNumber** field entries through the CAS check-digit verification process.[10] Entries that passed the mathematical verification process were flagged preliminarily as *Valid* (i.e., mathematically valid), subject to further validation against actual ingredient databases as discussed below. Entries that did not pass the mathematical verification process at this stage were labeled *Invalid*.

We noted the presence of "leading zeroes" preceding several of the *CAS number* values in the data set. We confirmed mathematically the redundancy of leading zeroes in the CAS verification algorithm. They do not change the identity of the corresponding chemical ingredient, but can interfere with the database cross-validation process. In addition, two otherwise identical CAS numbers with varying occurrences of leading zeroes would be treated as two separate CAS numbers. We therefore created a new **CASNumberClean** field by removing leading zeros from mathematically valid CAS numbers. All otherwise labeled **CASNumber** entries were listed as *NA* in this new field as they were not relevant to the subsequent analysis associated with the **CASNumberClean** field.

---

[10] A correctly formatted CAS number has three sets of numeric values separated by hyphens. The first group contains 2-7 digits, the second group 2 digits, and the final is a single digit that serves the purpose of a check digit. We went to the Chemical Abstract Service website https://www.cas.org/content/chemical-substances/checkdig and followed their instructions as below for mathematically verifying CAS numbers:

> The CAS Registry Number may be written in general form as:
>
> $N_i.....N_4N_3 - N_2N_1 - R$
>
> …in which R represents the check digit and N represents as fundamental sequential number. The check digit is derived from the following formula
>
> $(iN_i + ... + 4N_4 + 3N_3 + 2N_2 + 1N_1) / 10 = Q + R/10$
>
> ...in which Q represents an integer to be discarded.
>
> R must then match the original check digit for the CAS Number to be valid.

Mathematical validation only says that a CAS number's format is valid. Its reference to an actual chemical would need to be tested against databases, which was undertaken as described in Section 4.3.3.

4.3.3. CAS Number Standardization, Part III: Database Validation and Standardization of Ingredient Names

The next step was to verify that the mathematically *Valid* CAS numbers actually referred to a chemical by checking against online databases.[11] The database validation process was conducted on a newly created **CASNumberClean** field; entries in this field corresponding to a **CASLabel** entry other than *Valid* were listed as *NA* and not included in the database ingredient validation process. The original dataset contained 1000 unique and mathematically valid CAS numbers. 995 of the original 1000 unique CAS numbers were successfully run through the database reference process, and referred to a total of 981 unique ingredients.[12] **CASNumber** field entries corresponding to the five CAS numbers that were not verifiable against the seven databases were re-flagged as *Invalid* in the **CASLabel** field.

The database validation step provided an objective method for standardizing entries in the **IngredientName** field. We created a new **IngredientNameClean** field which we populated with the standardized name obtained from referencing the CAS databases. The NIH website lists the main name for "Ethyl Alcohol" as simply "Alcohol"; we manually standardized this to "Ethyl Alcohol" as a more chemically valid descriptor. This was the only instance of manual standardization.

For purposes of our subsequent analyses, we assume that the valid *CASNumber* values were entered correctly and refer to the chemical additive used in a particular job. Future work could involve an in-depth analysis quantifying instances where the **IngredientName** and **IngredientNameClean** fields refer to different chemicals.

4.3.4 Fields Relating to Disclosure of Chemical Information

The criteria for a *CASNumber* value being labeled as *Valid* or *Invalid* in the **CASLabel** field were discussed in Sections 4.3.2 and 4.3.3. *Trade Secret*, *Proprietary*, or *Confidential* values were entered into the **CASLabel** field based on exact phrase searches as shown in Table 2. We then created a logical **WithheldFlag** field, flagged as *TRUE* if a *CASLabel* value was *Confidential, Proprietary, Trade Secret*, or *N/A*, and *FALSE* if otherwise. We anticipated using this field to calculate rates of data withholding across all records and by justification.

---

[11] The databases were consulted in the following order, based on ease of access and comprehensiveness. For instance, SciFinder, despite being the most comprehensive, requires manual entry of each CAS number unless the user has a subscripton. Therefore, we used this database last, to look up the remaining 27 unverified CAS numbers.

NIH:	http://chem.sis.nlm.nih.gov/chemidplus/rn/
EPA:	http://ofmpub.epa.gov/sor_internet/registry/substreg/searchandretrieve/substancesearch/search.do?
ChemNet:	http://www.chemnet.com/
NIST:	http://webbook.nist.gov/chemistry/cas-ser.html
CommonChemistry:	http://www.commonchemistry.org/
SigmaAldrich:	http://www.sigmaaldrich.com/catalog/AdvancedSearchPage.do
SciFinder:	https://scifinder.cas.org/

[12] Multiple CAS numbers may refer to a single ingredient when a CAS number is replaced and users continue to report the retired CAS number. Fourteen chemicals had two valid CAS numbers associated with them.

We also anticipated analyzing the data to determine whether an emerging method of chemical reporting reduces the rate of chemical record withholding. The "systems approach" method aggregates all of the ingredients in the fracturing fluid rather than associating each ingredient with a particular product added to the fluid. When key words or phrases reflecting use of the systems approach appeared in the **CASNumber** or **IngredientName** fields, as shown in Table 2, we entered *Systems Approach* in the **CASLabel** field.

Finally, we added a separate and comprehensive **Systems_Approach** logical field (Table 3). We used this field to designate entries where major suppliers (see Section 4.2.5) used the "systems approach" method of disclosure. The **Systems_Approach** field was flagged *TRUE* if it was flagged as *Systems Approach* in the **CASLabel** field and/or if multiple product names were entered into a single **TradeName** record. All entries in the **Systems_Approach** field corresponding to suppliers other than the four major suppliers were flagged as *NA* as we did not rely on them for our subsequent analyses.

5. Conclusion

This paper documents the steps taken to develop a clean and standardized data set from forms submitted to FracFocus through April 13, 2015. Our data set, after passing forms through two primary QA criteria, contains data from 43,366 FF 1.0 forms, and 53,083 FF 2.0 forms. After applying a secondary criterion relating to state location, we had an analysis subset comprised of data from 43,222 FF 1.0 forms, and 53,078 FF 2.0 forms. At the L4 level of data, the forms describe 96,300 hydraulic fracturing jobs that took place at 92,844 wells in twenty-three states, and contain up to 17 additional fields (11 in the header data and 6 additional in the FF 2.0 dataset). The final set of new fields created are **search_id_form**, **submit_time_fracday**, **LatitudeClean**, **LongitudeClean**, **ProjectionClean**, **StateNameFromCoords**, **StateAbbFromCoords**, **CountyNameFromCoords**, **CountryCodeFromCoords**, **StateNameFromAPI**, **StateOK**, **SupplierClean**, **CASNumberClean**, **CASLabel**, **WithheldFlag**, **Systems_Approach**, and **IngredientNameClean**. We used the L4 data set to conduct all further analyses. All data (L1 through L4) may be found at http://dx.doi.org/10.7910/DVN/EFNV5J and is free and available for public use.

Table 2. CASLabel values.

User-input CASNumber values	CASLabel
"TradeSecret", "TRADESECRET", "Tradesecret", "TradeSeceret", "TradeSecret,disc.", "tradesecret" , "TRADESECRETS", "TradeName", "TradeSecert", "tracesecret", "tradeSecret", "TRADESECERET", "TradeSecrte", "TradeSecrer", "tradeseccret", "TradSecret", "ts", "TS"	*Trade Secret*
"Proprietary","PROPRIETARY","3rdPartyProprietar", "proprietary","Propietary","Prop","ProprietaryBlend", "Propreitary", "PROPRITARY", "PRORIETARY", "Proprietaryl" , "Proptietary", "PROP", "Proprietarty", "Prop.", "Proprietory", "Propreitory", "PROPRIETARY0.10" "proprietry", "propietary", "Proprietatry", "Propriatary" "propriatary", "Propritary", "Proprietar" "PROPRIERTARY", "Priprietary", "7732-18-5/propr", "Proprietart", "Proprieatary", "Proprietery", "PRIOPRIETARY", "PROPRIERARY", "Propriety", "prop", "proprietarty", "7732-18-5proprietary", "propriety", "Properitary"	*Proprietary*
"Confidential", "ConfidentialBusines", "Confidnetial", "Confidentail", "confidential","CONFIDENTIAL", "CONFIDENTIALBUSINES", "BusinessConfidental", "ConfidentialInfo","COnfidential", "Confidental", "Conf", "ConfBusInfo", "Confid.Bus.Info", "Confidenial", "Confidentia1" , "ConfidentialInf", "Confinential", "Confindential", "CBI"	*Confidential*
"N/A", "NotListed", "NONE", "none", "NotAvailable", "N.A.", "NOTPROVIDED", "None", "n/a", "NOTASSIGNED", "\"N/A\"", "-", na", "Notapplicable.", "Notavailable.", "N/D", "NotEstablished", "Notavailable", "NOTAVAILABLE", "NULL", "notlisted", "NA", "NDA", "Notlisted", "N\\A", "non", "NotApplicable", "N/a", "Undisclosed", "UNK", "unknown", "Unknown", "undisclosed", "unk", "Unavailable", " "	*Not Available*

Table 3. Systems Approach Values

User-input value designating systems approach	Search Criterion	Search Field
"Listed with chemicals", "Listed with Ingredients","Listed with Other Chemicals", "Listed with Other ingredients", "Listed with Chemicals", "Listed with Chemical Ingredients"	Exact Phrase	**IngredientName**
"Listed Below"	Keyword	**IngredientName**
"Listed", "SystemDisclosure"	Exact Phrase	**CASNumber**