Detailed Data Cleaning and Preparation Methods for
Public Reporting of Chemicals in Hydraulic Fracture Fluids: Withholding Rates after Disclosure Policy Change in FracFocus 3.0 (P. Sanghavi, K. Trickey, N. Hadjimichael)

March 12, 2019
Kevin Trickey

Introduction and Purpose

This paper describes the process of cleaning and preparing the data for the paper *Public Reporting of Chemicals in Hydraulic Fracture Fluids: Withholding Rates after Disclosure Policy Change in FracFocus 3.0*. Datasets from each level of cleaning, described below, are available for public use at https://drive.google.com/drive/folders/1aFHUQh9AthIOrQ4QlYKRRLJJhPTniCeU?usp=sharing; a repository of our cleaning and analysis code is available at https://gitlab.com/uchicago-fracking/fracfocus-analysis. A detailed, bulleted list of preparation steps and exact values of dataset modifications is contained in the file "2 Cleaning/Data Cleaning Log.txt" in our repository. Data preparation, including most aspects of L1–L4 cleaning, was modeled and extended from "FracFocus Chemical Disclosure Registry 1.0 and 2.0 Data Conversion and Cleaning Methods Paper" by Archana Dayalu and Kate Konschnik, whose dataset and methods paper is available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EFNV5J.

FracFocus (FF) is an online chemical disclosure registry for hydraulic fracturing operations, designed to make chemical compositions of fracturing fluids used in the United States more publicly accessible. At the time of our analysis, the entire FF registry is downloadable from https://fracfocus.org/data-download as a Microsoft SQL Server 2012 database backup, updated on the first and fifteenth of every month. The same data is available in comma-separated values (CSV) format. FF also contains a Find a Well utility, which allows the public to search for individual wells or drilling operations.

Data cleaning was split into five levels (L1–L5). All procedures were performed on an initial download from FF on January 26, 2018. As a one-year update, the FF database was downloaded again on January 19, 2019, and all five levels of preparation were performed on the new data. Throughout this paper, numbers describing the original download will appear in plain text, while numbers in brackets refer to the 2019 update. Because of the way the update file was processed, numbers referring to exclusions from the 2018 and 2019 downloads may not necessarily be mutually exclusive.

L1 and L2 Cleaning: Basic Formatting

L1 cleaning consisted of aggregating FracFocus downloads into one master table. Data was downloaded in a compressed CSV format, with 13 [17] separate files containing a maximum of 250,000 rows each. Each row represents one raw chemical, belonging to one of several fracture fluid ingredients that comprise one drill job.

L2 cleaning involved searching for any non-ASCII characters present in any character field in the data download. 9,805 [1,023] such characters, all of which were the newline "\n" or tab "\t" characters, were found and deleted.

L3 Cleaning: Standardization of Data Entry Spellings and Formats

Because our analysis centered on the transition from FF version 2.0 to 3.0, we excluded all 43,972 [88,184] rows submitted using FF 1.0. Inspection revealed that 1,428 [351] job submissions were duplicate submissions under different upload key values; these were excluded. As a further quality control step, 983 [2,434] rows indicating job start dates before January 1, 2011 were excluded.

To ensure consistency, we added a binary flag indicating whether the reported latitude and longitude of each well matched the reported county and state identifiers. Coordinates were reported in one of three projections (NAD27, NAD83, WGS84). Because NAD27 coordinates may differ from the other two by tens or hundreds of meters, 57,383 [10,190] wells using the NAD27 projection were converted to NAD83 coordinates using the NGS Coordinate Conversion and Transformation Tool (NCAT) Multipoint conversion tool (https://www.ngs.noaa.gov/NCAT/). Latitude/longitude points falling outside the range of the United States were inspected for obvious errors, such as missing negative signs or decimal points, and corrected where applicable. Converted and corrected values were assigned to new columns, **LatitudeClean** and **LongitudeClean**. Then, we fed the cleaned coordinates into the FCC Area API (https://geo.fcc.gov/api/census) to assign new columns **StateNameFromCoords**, **StateNumberFromCoords**, **StateAbbFromCoords**, **CountyNameFromCoords**, and **CountyCodeFromCoords**. Two more logical columns, **StateOK** and **CountyOK**, were assigned "True" if and only if these fields matched the reported state/county of the fracturing wellhead. In total, less than 0.1% of forms did not have matching state values, and 13% of forms did not have matching county values.

Alternate or mistaken spellings of company names were also a large point of discrepancy across form submissions. Using the **Supplier** field of string values, we generated the file "supplier aliases.csv", containing for each company a list of different strings that refer to it. Different values appearing in the data were first assigned automatically using Levenshtein string distances to the extent possible, then the matching was confirmed and completed manually. Each **Supplier** string maps to one standardized company name, or to the string "UNIDENTIFIABLE" if no such company was able to be determined. In cases where more than one supplier was listed for the same chemical, the first was used. Standardized names were appended to the dataset under a new **SupplierClean** column.

L4 Cleaning: Determining Withholding

Due to the high variability in chemical names and spellings, chemical information was determined using the **CASNumber** field, which carries user-supplied Chemical Abstracts Service (CAS) numbers identifying raw fracturing fluid chemicals. We created a **CASLabel** field carrying one of seven values, according to Table 1. To achieve a *Valid* label, a **CASNumber** first needed to satisfy the CAS Check Digit Verification process.[1] Following this, it was referenced against seven different chemical databases, including the CAS Scifinder® tool.[2] Only CAS numbers confirmed in at least one database were marked

[1] https://www.cas.org/support/documentation/chemical-substances/checkdig. Valid CAS numbers will satisfy a mathematical Check Digit Verification process, as follows. The CAS number may be written as $N_i \cdots N_4 N_3 - N_2 N_1 - R$, where R is the "check digit." The digits must satisfy the equation $\frac{1}{10}(iN_i + \cdots + 4N_4 + 3N_3 + 2N_2 + 1N_1) = Q + \frac{R}{10}$ for some integer Q.

[2] Databases were consulted in the following order, based loosely on completeness and ease of automation:

1. NIH: http://chem.sis.nlm.nih.gov/chemidplus/rn/ (1,129 [59] names verified)
2. EPA: http://ofmpub.epa.gov/sor_internet/registry/substreg/searchandretrieve/substancesearch/search.do? (manual - 1 name verified)

Valid; a new column **CASNumberClean** was created containing *NA* for non-valid rows and clean CAS numbers (i.e., leading zeros trimmed) for valid ones. Furthermore, all valid chemicals were assigned a standardized chemical name, based on these databases, in a new **IngredientNameClean** column.

Further investigation revealed that many *Invalid* entries were in fact disclosures of water (or saltwater) where users had not included its CAS number. To correct for this, the **IngredientName** field was searched for names referring to water, and the appropriate values were assigned to **CASNumberClean** and **IngredientNameClean**. In addition, the **CASLabel** for every such entry was reassigned as *Valid*.

Finally, forms that used the systems approach contained separate rows to hold ingredient trade names and identifying CAS numbers. To ensure that rows of the former type were not falsely labeled as *Invalid*, the data was exhaustively inspected for indications of a systems approach form (Table 1). All data rows containing trade names but using systems-disclosed chemical identifiers were given a **CASLabel** of *Systems Approach*; these were excluded from our final analysis, since we were concerned with chemical compositions rather than ingredient trade names. Because the systems approach was used at the form rather than the ingredient level, a new Boolean field **FormUsedSystemsApproach** was created to indicate for each row whether any chemical on the same form had been labeled *Systems Approach*.

A **WithheldFlag** column was then added to the data, containing *True* if and only if the **CASLabel** was *Valid* or *Systems Approach*.

L5 Cleaning: Gathering of Full Data

In the initial 2018 download, systems approach forms from FF 3.0 only contained information on trade names and excluded the aggregated list of chemical identifiers, leaving crucial gaps in our dataset. To fill these, each systems approach form in question was searched in FF's "Find a Well" utility (https://fracfocusdata.org/DisclosureSearch/Search.aspx); then, the appropriate PDF disclosure file was downloaded and the chemical data extracted. Automation of these processes was performed in Python 3.6 using Selenium WebDriver and tabula-py. Extracted data was joined with the original set, and a new Boolean column **FromPDF** was added to flag the rows added. By the 2019 download from FF, chemical information was no longer missing from the downloaded set and this process was unnecessary.

Finally, a **DownloadDate** field was appended to track each form's original dataset, either *2018-01-26* or *2019-01-19*.

3. ChemNet: http://www.chemnet.com/ (9 [0] names verified)
4. NIST: http://webbook.nist.gov/chemistry/cas-ser.html (5 [0] names verified)
5. CommonChemistry: http://www.commonchemistry.org/ (0 [0] names verified)
6. SigmaAldrich: http://www.sigmaaldrich.com/catalog/AdvancedSearchPage.do (2 [0] names verified)
7. SciFinder: https://scifinder.cas.org/ (manual - 46 [7] names verified)

***CASLabel* value assigned**	**Criteria**
Valid	Mathematically valid **CASNumber** after removing any leading zeros and confirmed by an online chemical database, OR **IngredientName** one of the following strings: "Water (Including Mix Water Supplied by Client)*", "Water", "Carrier / Base Fluid - Water", "2% KCL Water", "Fresh Water", "Water (Including Mix Water Supplied by Client)", "Water, other", "Recycled Water ", "4% KCL Water", "Brine Water", "Lease Water", "NFIDB:2% KCL Water", "3% KCL Water", "Field Salt Water", "Produced Brine Water", "Tulare Water", "NFIDB:Lease Water", "water", "Water ", "KCl Water", "Produced Water", "1% KCL Water", "NFIDB:4% KCL Water", "NFIDB:Brine Water", "Recycled Water", "4% Salt Water", "10% Salt Water", "2% KCl Water", "NFIDB:3% KCL Water", "Water Moisture", "6% KCL Water", "fresh water", "NFIDB:3% NaCl Water", "NFIDB:6% KCL Water", "NFIDB:7% KCL Water", "18% Salt Water", "3% NaCl Water", "3% NACL Water", "4% NaCl Water", "5% KCl Water", "5% KCL Water", "7% KCL Water", "water, other", "2% KCl Lokern Water", "Brackish Water", "Brackish Water ", "NaCl Water", "NFIDB:15% Salt Water", "15% Salt Water", "3% Salt Water", "Fresh water", "KCL Water", "Lease Salt Water", "lease water", "NFIDB:5% KCL Water", "NFIDB:Water", "Produced Water ", "Production Water", "Salt Water", "Water (including mix water supplied by client)*", "Water, Other", " Water ", "1.5% KCL Water", "10% KCL Water", "3% KCl Water", "4% KCl Water", "4% NaCl Water ", "6% Salt Water ?", "Field Water", "NFIDB - 4 percent KCL Water", "NFIDB:1% KCL Water", "NFIDB:10% KCL Water", "NFIDB:10% Salt Water", "NFIDB:Sea Water", "produced Water", "Seawater", "Tulare Water", "Water (Including Mix Water Supplied by Client).", "Water (including mix water supplied by Client)*", "Water (including Mix Water supplied by Client)*", "Water (major)", "Water, Including MIx water supplied by client", "Water,other", "water(including mix water supplied by client)*", "Water/Salt"
Proprietary	**CASNumber** one of the following strings: "PROPRIETARY", "Proprietary", "proprietary", "Proprietatry", "3rdPartyProprietar", "Proprietar", "ProprietaryBlend", "Prop", "7732-18-5proprietary", "PROP", "7732-18-5/propr", "Propritary", "PROPRITARY", "propietary", "propriety", "3rdpartyproprietar", "Propietary", "proprietry", "Proprietart", "Priprietary", "Proprietery", "Properitary", "Propreitary", "Propriatary", "Prop.", "Proprietarty", "PROPRIERTARY", "PROPRIERARY", "PRIOPRIETARY", "prop", "Propriety", "proprietarty", "PRORIETARY", "Proprietaryl", "PROPRIETARY0.10", "Proptietary", "Proprietory", "Propreitory", "Proprieatary", "propriatary", "proprietory", "Porprietary"
Confidential	**CASNumber** one of the following strings: "Confidential", "CBI", "Confidnetial", "confidential", "CONFIDENTIAL", "ConfidentialInfo", "ConfidentialBusines", "Confidenial", "Confidentail", "ConfBusInfo", "Confid.Bus.Info", "Confidental", "BusinessConfidental", "Conf", "Confidentia1", "Confinential", "CONFIDENTIALBUSINES", "Confindential", "COnfidential", "Condidential", "Confdential", "Coinfidential"
Trade Secret	**CASNumber** one of the following strings: "TradeSecret", "TRADESECRET", "tradesecret", "TradeSecret,disc.", "Tradesecret", "TRADESECRETS", "TS", "ts", "tradeseccret", "TradeSeceret", "tracesecret", "TradeName", "TradSecret",

	"TRADESECERET", "tradeSecret", "TradeSecrer", "TradeSecrte", "TradeSecert", "Tradesecret.", "Trade"
Not Available	**CASNumber** one of the following strings: "NOTPROVIDED", "N/A", "na", "NA", "NotAvailable", "N.A.", "None", "none", "NOTAVAILABLE", "NONE", "n/a", "unknown", "Unavailable", "Notavailable", "Notavailable.", "\"N/A\"", "NOTASSIGNED", "undisclosed", "NotApplicable", "notlisted", "Undisclosed", "notassigned", "Notapplicable.", "UNK", "N/D", "N\\A", "NULL", "unk", "Notlisted", "NotEstablished", "non", "n/A", "Unknown", "NotAssigned", "NA?", "N/a", "Na", "(N/A#)", "BA", "Blank", "CASNotAssigned", "NoneListed", "UNKNOWN", "Notassigned", "NotListed", "CASnotassigned"
Systems Approach	**CASNumber** one of the following strings: "Listed", "SystemDisclosure", "ListedBelow", OR **IngredientName** one of the following strings: "Listed with chemicals", "Listed with Ingredients", "Listed with Other Chemicals", "Listed with Other ingredients", "Listed with Chemicals", "Listed with Chemical Ingredients", OR **IngredientName** contains the substring "listed below", ignoring letter case
Invalid	Data row did not match any of the above criteria

Table 1: Criteria for ***CASLabel*** *assignments. String criteria for* Proprietary, Confidential, Trade Secret, *and* Not Available *were determined by an exhaustive inspection of the data. Other than these four reasons, no other justifications were found for withholding.*