Facility Registry Service Best Pick Processing

Delivery Order # GS00Q09BGD0022
Task Order # 47QFCA-18-F-0009

Version 1.0
April 16, 2019

TABLE OF CONTENTS

Exhibits iv
1 Introduction to the Facility Registry Service 6
1.1 FRS Registry and Program Records 6
1.2 Purpose of this document 7
2 FRS Best Pick 7
2.1 Geocoding 7
2.2 Quality Assurance Check 9
2.2.1 Accuracy Value 11
2.3 Best Pick Processing 13

EXHIBITS

Exhibit 1-1: Development of FRS Registry Records from EPA Program Facility Records, Partner Facility Records, and FLA Use 6

Exhibit 2-1: Schematic of FRS Geocoding. 8

Exhibit 2-2: Collection Method Codes 9

Exhibit 2-3: Reference Point Codes 10

Exhibit 2-4: Maximum NAD27 Datum Shift by State 11

Exhibit 2-5: FRS Best Pick Process 15

REVISION LOG

Date	Version No.	Description	Author	Reviewer	Review Date
04/16/2019	V 1.0	Draft version of Facility Registry Service Best Pick	K. Hoke	M. Kelly	4/15/19

1 INTRODUCTION TO THE FACILITY REGISTRY SERVICE

Facility data is at the core of federal, state, local, and tribal (SLT) environmental regulatory processes. Knowing a facility's name, ownership, location, and characteristics are key to a comprehensive picture of past, current, future, and potential environmental impacts. Linked to other critical environmental data such as ambient air and water quality data, census figures, and other demographic information, facility data has the capacity to provide a broad perspective of the facility, enabling co-regulators to better protect human health and the environment.

The Facility Registry Service (FRS) is Environmental Protection Agency's (EPA's) source for integrated facility information, managing integrated data for ***more than 5.2 million facilities***. This information is provided for public viewing via queries and in other EPA platforms such as Envirofacts, Cleanups in My Community (CIMC), and Enforcement and Compliance History Online (ECHO).

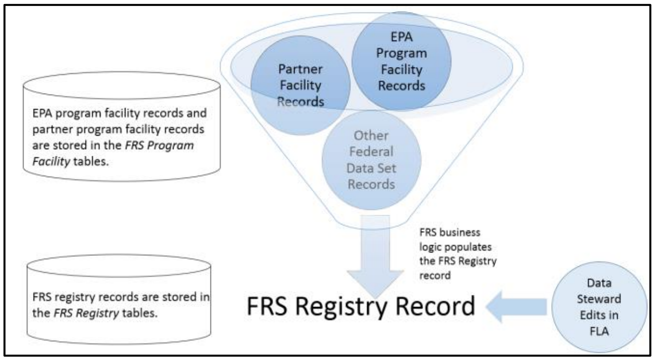
FRS links facility, geospatial, and reference data across EPA programs, states, tribes, and other federal agencies based on a consistent set of data elements such as address, facility name, and location. The main goals of FRS are to integrate data, improve data, and provide data to EPA programs, partners, and the public. FRS allows users to identify the permits or regulations that apply to a particular facility, the location of a facility, and regulated facilities within a specific sector. FRS is used for program management, enforcement, rulemaking, reporting, analysis, and emergency response.

By integrating disparate sources of facility information together, FRS provides critical comprehensive information about a facility's regulatory processes and environmental impact. FRS also enhances facility data by applying data validation and mapping capabilities, further enabling the use of this critical data.

1.1 FRS REGISTRY AND PROGRAM RECORDS

FRS links facility data from a variety of sources into a single record, called the FRS registry record. The FRS registry record is linked to one or more program records provided by EPA program systems and partners (states, tribes, and local agencies). Each program record contains its own attributes for facility name, address, and location coordinates, among others. FRS uses a set of processes to determine how to populate the FRS registry record. FRS also has a tool, the Facility Linkage Application (FLA), which allows FRS data stewards to make updates to a FRS registry record. *Exhibit 1-1* illustrates the creation of FRS registry records from EPA program facility records, partner facility records, and through the use of the FLA.

Exhibit 1-1: Development of FRS Registry Records from EPA Program Facility Records, Partner Facility Records, and FLA Use


EPA Program Facility Records
Partner Facility Records
Other Federal Data Set Records
EPA program facility records and partner program facility records are stored in the *FRS Program Facility* tables.
FRS business logic populates the FRS Registry record
FRS registry records are stored in the *FRS Registry* tables.
FRS Registry Record
Data Steward Edits in FLA

1.2 Purpose of this Document

This document describes the logic used in FRS to determine the Best Pick location to represent a registry-level facility record. Although FRS is in the 'Operation and Maintenance' phase of development, there are ongoing changes and improvements to FRS, including potential changes to the Best Pick process. This document will be updated as those changes are implemented.

2 FRS Best Pick

The FRS best pick represents the location (latitude/longitude) that best represents the physical location of the FRS registry record. The best pick process described in this document populates the latitude, longitude, and metadata attributes in the FRS registry record. The FRS best pick is identified as the registry facility location within the public and internal FRS databases. The best pick process is run once and updates the internal (non-public) version first. Then, public data is pushed from the internal version to the external version.

The FRS best pick is determined by evaluating the program records linked to a FRS registry record. The factors used to determine the FRS best pick include evaluating:

- The program records associated to the FRS registry record.
- The address and location coordinate information in the program records.
- The date the program records were most recently updated.

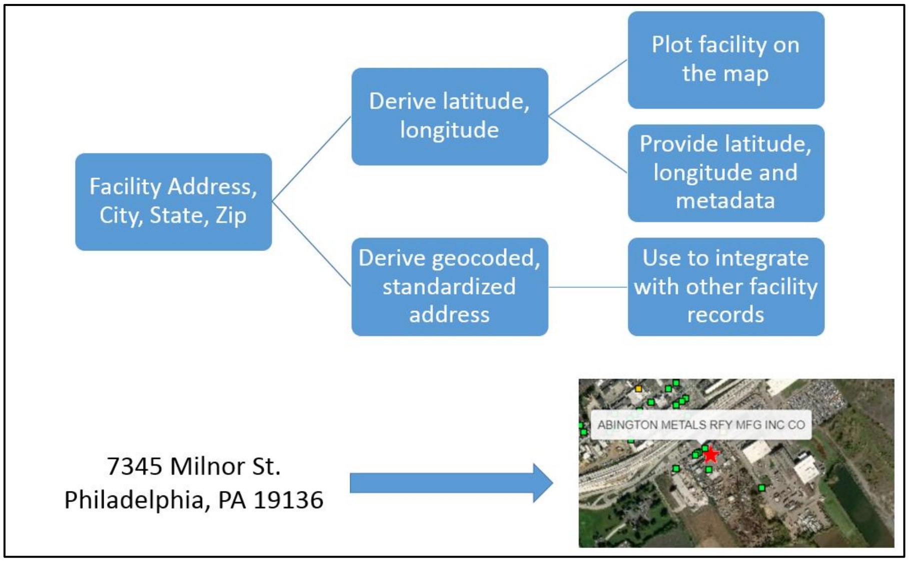
It is possible that a FRS registry record will have no best pick value. This occurs if the conditions described in this section are not met (i.e., if the process cannot complete for a given record). FRS runs the best pick process every weekend. Geocoding and the Quality Assurance (QA) check support the best pick process. See ***Sections 2.1*** and ***2.2*** for descriptions of those processes.

The FRS Best Pick location is used in other geospatial processing in FRS and is also provided and displayed on maps and in search results for many EPA searches and services.

2.1 Geocoding

Geocoding is used to translate a physical address, such as a street address, to a set of coordinates (latitude and longitude values) and is performed after source program data is loaded into staging tables in FRS. The geocoding process is illustrated in *Exhibit 2-1*.Geocoding uses a piece of software (geocode) that evaluates physical addresses and determines the appropriate latitude and longitude values that correspond to that address. The ability to translate a street address into a point that can be plotted on a map is especially important when location (latitude/longitude) data was not widely available for most data sets.

Exhibit 2-1: Schematic of FRS Geocoding.


Facility Address, City, State, Zip
Derive latitude, longitude
Plot facility on the map
Provide latitude, longitude and metadata
Derive geocoded, standardized address
Use to integrate with other facility records
7345 Milnor St.
Philadelphia, PA 19136
ABINGTON METALS RFY MFG INC CO

The geocoder standardizes FRS program facility spatial coordinates to the United States Standard North American Datum of 1983 (NAD83) horizontal datum using an Oracle Spatial geocoder routines along with HERE Navteq Point Address data. The HERE data is refreshed quarterly and historically, the FRS team has updated this dataset within FRS annually. The most recent update was in October 2018. The update begins with the National Computer Center (NCC) loading the latest HERE data into a local database. FRS then pulls the latest dataset and re-geocodes all of the data in FRS.

Using the Oracle Spatial Geocoder, all new FRS facilities, and all existing FRS facilities for which any part of the address has been modified or edited, are geocoded. The geocoding process follows these steps:

1) Modify facility address, replacing certain abbreviations and address components with values that the Oracle geocoder can more readily recognize. Examples include replacing “TWP” with “TOWNSHIP OF,” and “BORO” with “BOROUGH OF.”
2) Identify each intersecting street of intersections within the FRS location address.
3) Geocode a formatted address using the Oracle Spatial geocoder.
4) Validate the returned geocode object from the geocoder. That validation follows the steps described below:
    a) Reject returned geocodes that do not meet FRS accuracy standards and accept only reference point (e.g., rooftop), house number, street, intersection, and landmark place name geocodes. Geocoded data is given an accuracy score based on a reference point such as a house number, street, etc. The more specific the reference point, the better. For example, a house number reference point is better than a street reference point. If the geocoded address does not receive an accuracy score, then it is rejected. The cause of this is usually incomplete or missing data.
    b) For reference point and house number geocodes, validate that the returned house number matches the supplied house number. If not, convert the geocode to the “Street” geocode.
    c) Using the Jaro-Winkler and Levenshtein Distance algorithms, validate the returned street name/place name against the supplied street name/place name. If they do not match, reject the geocode.

d) For street level geocodes, identify the number of blocks for the street and the length of the street. If the street is longer than two (2) kilometers, and if the street contains more than (two) 2 blocks, reject the geocode.

e) Validate the state returned from the geocoder. If it does not match the supplied state, reject the geocode.

f) Validate the Zone Improvement Plan (ZIP) code, county, and city returned from the geocode together using the United States Postal Service (USPS) dataset. This dataset was last updated in 2015. In certain conditions, Oracle will return the county name instead of the city name for rural areas. If the city, county, and ZIP code cannot be validated, reject the geocode.

Not all addresses can be geocoded. Some program records do not have a complete address, or have values of "unknown" for the address. In other cases, the geocoder does not recognize the address in the record. During geocoding, addresses are evaluated, and if they can be geocoded, they are given a score to indicate the level of confidence in the geocoded location. The score is stored with FRS in the "accuracy_value" column. The lower the accuracy value, the better the score. These levels of confidence and their corresponding accuracy values, in decreasing order, are:

1) Point (e.g., rooftop, center of facility) – 30 meters
2) Street House Number – 150 meters
3) Intersection – 200 meters
4) Street of only one 1 block – 500 meters
5) Street of only (two) 2 blocks or less than two (2) kilometers in length – 2,000 meters
6) Place name – 4,500 meters

2.2 Quality Assurance Check

The QA check process evaluates the facility coordinates (latitude and longitude) and associated metadata, provided in the program record, and determines whether it is within the ZIP code, city, county, and state boundaries provided in the program record. Not all program records provide latitude and longitude coordinates. Using program coordinate metadata for 'Collection Method,' 'Horizontal Datum,' 'Accuracy Value,' and 'Reference Point,' the FRS QA Check calculates an 'Accuracy Score' for all new and modified program coordinate records. These metadata elements are defined below:

Collection Method: The code that represents the method used to determine the latitude and longitude coordinates for a point on the Earth. See *Exhibit 2-2* for current Collection Method codes in FRS. Collection Method codes are based on Appendix B of the Latitude Longitude Data Standard.

Exhibit 2-2: Collection Method Codes

Collection Method Code	Description
001	ADDRESS MATCHING-HOUSE NUMBER
002	ADDRESS MATCHING-BLOCK FACE
003	ADDRESS MATCHING-CENTER LINE
004	ADDRESS MATCHING-NEAREST INTERSECTION

Collection Method Code	Description
007	ADDRESS MATCHING-OTHER
018	INTERPOLATION-MAP
019	INTERPOLATION-PHOTO
021	INTERPOLATION-OTHER
026	INACCURATE
027	UNKNOWN
028	GPS - UNSPECIFIED
039	GDT-ADDRESS MATCHING (GEOCODING)

Horizontal Datum: The code that represents the reference datum used in determining latitude and longitude coordinates.

Accuracy Value: The measure of the accuracy (in meters) of the latitude and longitude coordinates. ***Section 2.2.1*** discusses how FRS calculates the accuracy value.

Reference Point: The code that represents the place for which geographic coordinates were established. The code that represents the method used to determine the latitude and longitude coordinates for a point on the Earth. See *Exhibit 2-3* for current Reference Point codes in FRS. Reference Point codes are based on Appendix C of the EPA Latitude/Longitude Data Standard.

Exhibit 2-3: Reference Point Codes

Reference Point Code	Description
001	UNKNOWN
002	PLANT ENTRANCE (GENERAL)
003	OTHER
015	LOADING FACILITY
019	ADMINISTRATIVE BUILDING
020	FACILITY CENTROID
025	CENTER OF FACILITY
103	FACILITY/MONITORING SITE BOUNDARY POINT
006	AIR RELEASE STACK

Reference Point Code	Description
007	AIR RELEASE VENT
106	POINT WHERE SUBSTANCE IS RELEASED

2.2.1 Accuracy Value

The accuracy value is a score expressed in meters that estimates the accuracy of the collected coordinates. The lower the accuracy score, or modified accuracy score, the better, and more accurate the point is. FRS uses the following process to calculate the Accuracy Score:

1) If the collection method is missing, then set the accuracy score to 17,400. This represents the square root of the average area of all ZIP codes. A missing collection method defaults to the accuracy score of a ZIP code centroid. If the accuracy value is present, it is ignored because it cannot be validated against the collection method.
2) If the accuracy value is null, calculate the accuracy score based on the default accuracy for the collection method. If the accuracy value is not null, verify the maximum accuracy allowed for the collection method. If the reported accuracy value is greater than the maximum accuracy allowed for the collection method, then the accuracy score is equal to the accuracy value; otherwise, the accuracy score is equal to the default accuracy based on the collection method. This prevents situations where there is a reported accuracy value of “5 meters,” but the coordinates were obtained from map interpolation of a 1:100000 scale map, for example.
3) If the horizontal datum value is null and the collection method is one where North American Datum of 1927 (NAD27) could be used as the horizontal datum (e.g., map interpolation), then add to the accuracy score the maximum shift in meters between NAD27 and NAD83 coordinates, for the state where the coordinate is located. The datum shift values by state are illustrated in *Exhibit 2-4*, below.
4) Using the QA flags for the program coordinates, the accuracy score is adjusted if the program coordinates flunked any of the spatial boundary checks. QA flags indicate whether the program coordinates are within the ZIP Code, city, county, and state boundaries of the FRS registry record. The following values are potentially added to the Accuracy Score based on the QA check results:
   a) Flunk State Boundary: add 999999999.
   b) Flunk County Boundary: add 64228.
   c) Flunk City Boundary: add 18500.
   d) Flunk Zip Code Boundary: add 17400.
   e) If the reference point is null, or if the reference point is a facility boundary or a location on the street, add 30 to the accuracy score.

Exhibit 2-4: Maximum NAD27 Datum Shift by State

State	Maximum NAD27 Datum Shift
AK	214.91
AL	22.46
AR	23.1
AS	109.24

State	Maximum NAD27 Datum Shift
AZ	74.59
CA	100.9
CO	58.14
CT	42.24
DC	29.17
DE	34.93
FL	50.88
GA	31.44
GU	149.19
HI	466.15
IA	25.96
ID	80.87
IL	13.89
IN	8.52
KS	39.63
KY	17.88
LA	29.08
MA	47.3
MD	34.97
ME	46.47
MI	14.63
MN	27.69
MO	22.64
MS	22.47
MT	74.95
NC	41.65
ND	35.68

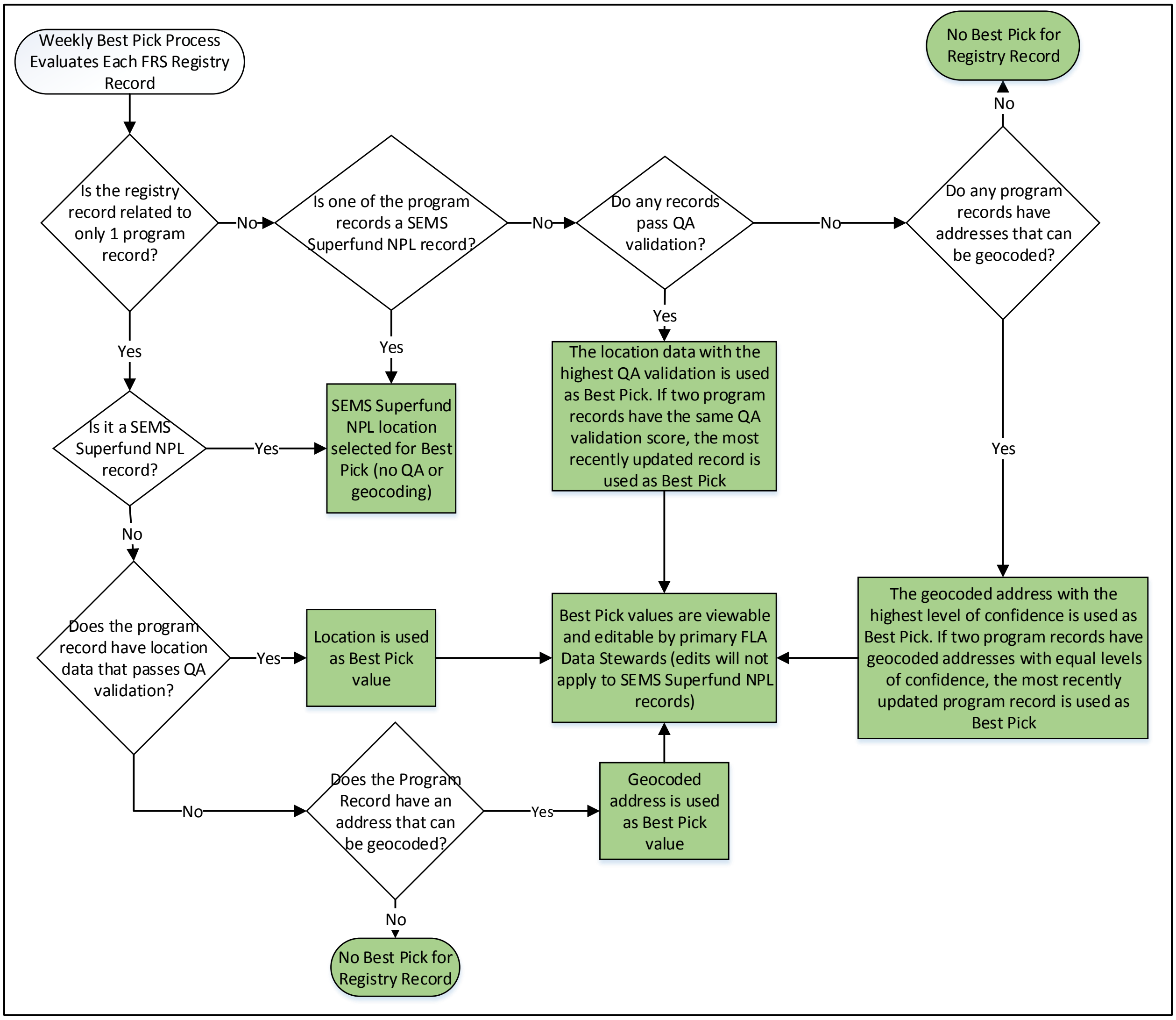
State	Maximum NAD27 Datum Shift
NE	42.37
NH	42.32
NJ	38.38
NM	58.61
NV	89.4
NY	42.48
OH	19.1
OK	41.44
OR	102.04
PA	35.81
PR	225.12
RI	44.17
SC	30.92
SD	38.81
TN	19.91
TX	52.99
UT	71.73
VA	35.24
VI	224.59
VT	39.31
WA	102.97
WI	16.85
WV	27.01
WY	63.2

2.3 Best Pick Processing

The FRS best pick process maintains a few baseline tables of best pick data. On a weekly schedule, the FRS best pick process evaluates incoming data (data added or updated in FRS) for inserts, updates, and deletes. In this way, it evaluates program records to make a determination of the best pick value (or does not assign a best pick value) based on the conditions described below and illustrated in *Exhibit 2-5*.

1) If there is a Superfund Enterprise Management System (SEMS) National Priorities List (NPL) program record related to the FRS facility record, the location coordinates within the SEMS Superfund NPL are selected for FRS best pick. FRS uses the 8R active list (available at: https://www.epa.gov/superfund/superfund-data-and-reports) to refresh this data. A SEMS Superfund NPL record is selected for best pick regardless of the outcome of geocoding the address and without performing a QA check on the coordinates. Best pick values populated from SEMS Superfund NPL program records cannot be overwritten by an FLA user.
2) If there is a single program record linked to the FRS facility record, and that program record is not a Superfund SEMS NPL record, then records that pass the QA check (see ***Section 2.2***) *or* records that can be geocoded can be used as best pick for the registry record. The best pick is selected based on the following conditions:
   a) If the address cannot be geocoded the program latitude and longitude values will be used for the best pick value.
   b) If the address can be geocoded, FRS performs a QA check on the location coordinates associated with the program record. See ***Section 2.2*** for an explanation of the QA check process.
      i. If the location coordinates pass the QA check, then the location coordinates become the best pick for the FRS facility record.
      ii. If the location coordinates do not pass QA validation nor show up within the newly modified accuracy score table, or if there is no location coordinates in the program record, the geocoded address becomes the best pick for the FRS facility record.
3) If there are multiple program records linked to the FRS facility record, none of which are SEMS Superfund NPL records, then FRS evaluates the program records to determine best pick as follows:
   a) FRS examines the program records to determine if any have passed the QA check and, if so, what their QA scores are. See ***Section 2.2*** for an explanation of the QA check process and scoring. If any program records pass the QA check, the location data with the lowest (i.e., best) QA score becomes the best pick record. If there is more than one record with the same QA score for location coordinates, then the record with the lowest modified accuracy score (see ***Section 2.2)*** becomes the best pick record.
   b) If no program location coordinates pass the QA check, then FRS examines the program records to determine the scores associated with geocoding the program record addresses. Refer to ***Section 2.1*** for an explanation of the geocoding process. The program record with the highest score for geocoding level of confidence is selected as the FRS best pick for the FRS registry record. If more than one program record has the same score for geocoding level of confidence, then the most recently updated program record's geocoded location becomes the FRS best pick for the FRS registry record. FRS determines "most recently updated" by using the date FRS integrated the record from the source system.
4) Once a best pick location is determined, it is part of the registry record that can be edited in FLA by data stewards. A FLA data steward may edit a best pick location by either moving the point on a map or by entering in location coordinate data. This will appear to change the best pick location within FLA; however, the new location is not identified as the best pick in the FRS registry record until the best pick process runs the following weekend. It should be noted that an edit made by a FLA data steward would not override the best pick location for those based on SEMS Superfund NPL records.

Exhibit 2-5: FRS Best Pick Process


Weekly Best Pick Process Evaluates Each FRS Registry Record
Is the registry record related to only 1 program record?
No
Is one of the program records a SEMS Superfund NPL record?
No
Do any records pass QA validation?
No
Do any program records have addresses that can be geocoded?
No
No Best Pick for Registry Record
Yes
Yes
Yes
Yes
The location data with the highest QA validation is used as Best Pick. If two program records have the same QA validation score, the most recently updated record is used as Best Pick
Is it a SEMS Superfund NPL record?
Yes
SEMS Superfund NPL location selected for Best Pick (no QA or geocoding)
No
The geocoded address with the highest level of confidence is used as Best Pick. If two program records have geocoded addresses with equal levels of confidence, the most recently updated program record is used as Best Pick
Does the program record have location data that passes QA validation?
Yes
Location is used as Best Pick value
Best Pick values are viewable and editable by primary FLA Data Stewards (edits will not apply to SEMS Superfund NPL records)
No
Does the Program Record have an address that can be geocoded?
Yes
Geocoded address is used as Best Pick value
No
No Best Pick for Registry Record