



Facility Registry System

Data Quality Enhancement, Integration and Corporate Parent To Support TRI Data and Analysis

November 3, 2010



FRS and TRI Data Quality

- Overview of FRS
- Data Quality Improvement Efforts



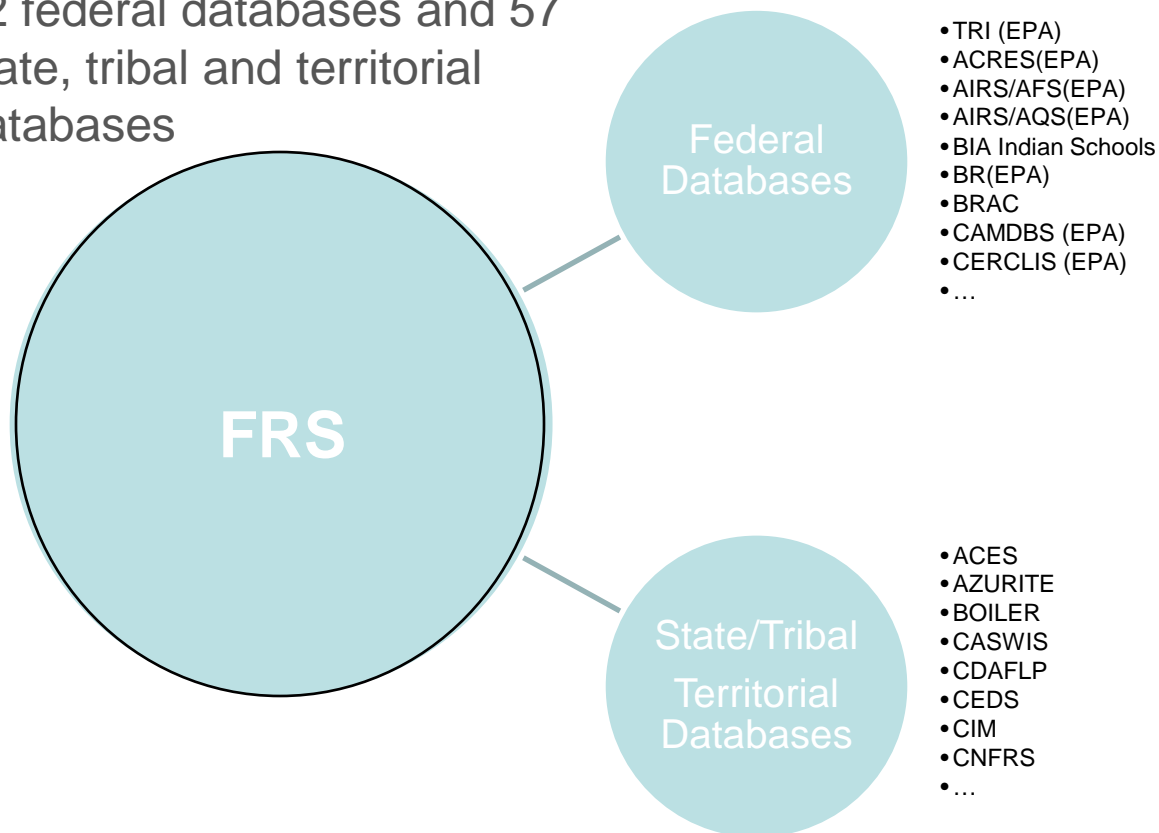
FRS and TRI Data Quality

- FRS Background
 - FRS Integrates data from across 32 Federal and 57 State, Tribal and Territorial Databases
 - Over 2 million facilities
 - Locational Data
 - Organizational Information – Corporate Parent
 - Program Identifiers, Permit Numbers
 - Points of Contact – Administrative, Operational and others
 - NAICS/SIC Codes



FRS and TRI Data Quality

Currently integrating data from 32 federal databases and 57 state, tribal and territorial databases



As of Nov 3, 2010



FRS and TRI Data Quality

- FRS aggregates and integrates, and in the process we attempt to use automation to validate addresses and other fields, for geocoding and other quality improvements
- Data quality varies...
 - Many fail validation on street address, city, county, ZIP codes and other fields
 - Data quality varies from program to program



FRS and TRI Data Quality

- Relationship between TRI and FRS
 - TRI performs facility lookups against FRS via Web Service
 - Existing, known facilities are mapped to, reducing duplication and redundancy, and improving information about the facility
 - Builds upon existing knowledge base for a given facility
 - Opportunity to concurrently improve data in TRI and FRS



FRS and TRI Data Quality

- Where does TRI stand?
- At present:
 - 54,826 total TRI facilities
 - 53,272 pass validation
 - 1,554 fail validation, typically one or more of invalid city, ZIP, county



FRS and TRI Data Quality

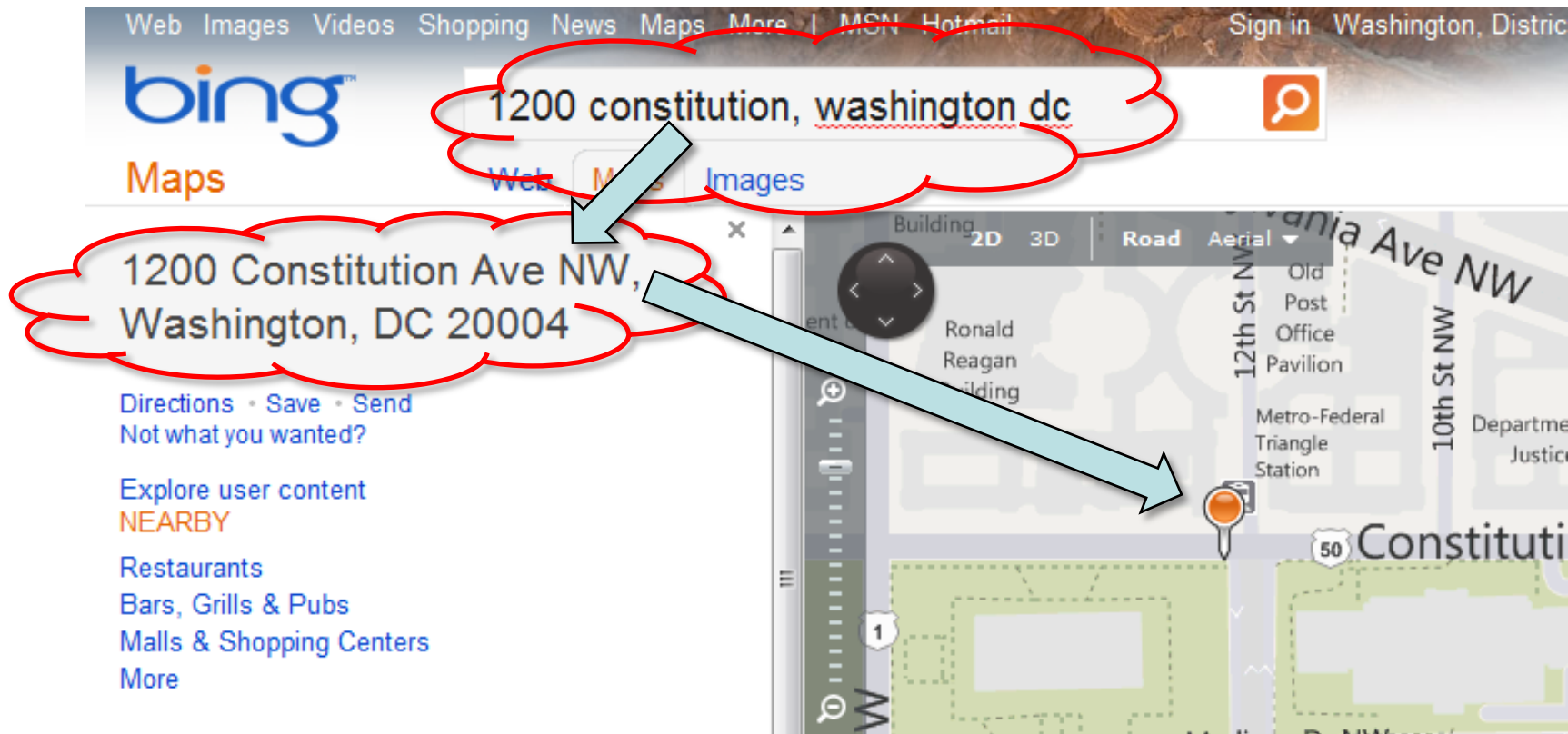
- What impact?
 - Ambiguity may lead to facility being identified in wrong community, may be miles from the actual location
 - Ambiguity may lead to mismatch of facilities, leading to duplicates or erroneous merges
 - Inability to geocode to a latitude/longitude to allow analysis of impacts at a community level



FRS and TRI Data Quality

- What can be explored?
 - Front-end locational data improvement – web services, address parsers, geocoders
 - Organizational Entity / Corporate Parent – web services, lookups, unique IDs
 - Crowdsourcing and community engagement – “Volunteered Geographic Information”, i.e. OpenStreetMap and others
 - Ubiquity of tools and technology to be leveraged, e.g. iPhone apps

Notional Ideas for Data Quality Improvement



Web services to validate, standardize, and improve data quality are becoming ubiquitous



FRS and TRI Data Quality

Notional Ideas for Data Quality Improvement



Dynamically-populated lists can be used to perform lookups, speed up and simplify data entry, and drive standardization in data entry



FRS and TRI Data Quality

- Locational Data Quality: Pushing the envelope
 - Typically 80-90% has been considered “good”: FRS is generally there now – TRI facilities validate to >95% but others are around 70% or less, FRS is currently at 82% overall, across all programs.
 - But is it good enough? Not for analysis and protecting communities
 - Many FRS facilities do not have standard street addresses
 - PLSS Section/Township/Range
 - “13.5 miles northwest of Elkhorn, Wyoming”
 - “Km 13.5 east of Guayama on Route 15”
 - Or... nothing at all
 - There are ways to handle these as well



FRS and TRI Data Quality

- Locational Data Improvement
 - Iterative development of location parsers and geocoders
 - Site Locator tool to allow users to locate their facility, drop a pushpin or trace a polygon
 - Potential crowdsourcing applications, like OpenStreetMap
 - Potential mobile apps



FRS and TRI Data Quality

- Organization/Corporate data improvement
 - Agency workgroup – TRI, TSCA, GHG, RMP and others
 - Leverage existing work on corporate name standardization in TRI program by ABT Associates
 - External workgroup – White House inter-agency effort to coordinate and collaborate on non-proprietary corporate/organization identifier: EPA, SEC, DOL and others
 - Corporate ID is far bigger than EPA alone, we need to collaborate across agencies



Contact:

David G. Smith PE PLS
US Environmental Protection Agency
Office of Environmental Information
Office of Information Collection
smith.davidg@epa.gov
202-566-0797