Data Dictionary/Harmonization

Combined Air Emissions Reporting (CAER)

Short-term Wins Project

John Harman and Lauren Gordon

US EPA Office of Environmental Information

Tammy Manning

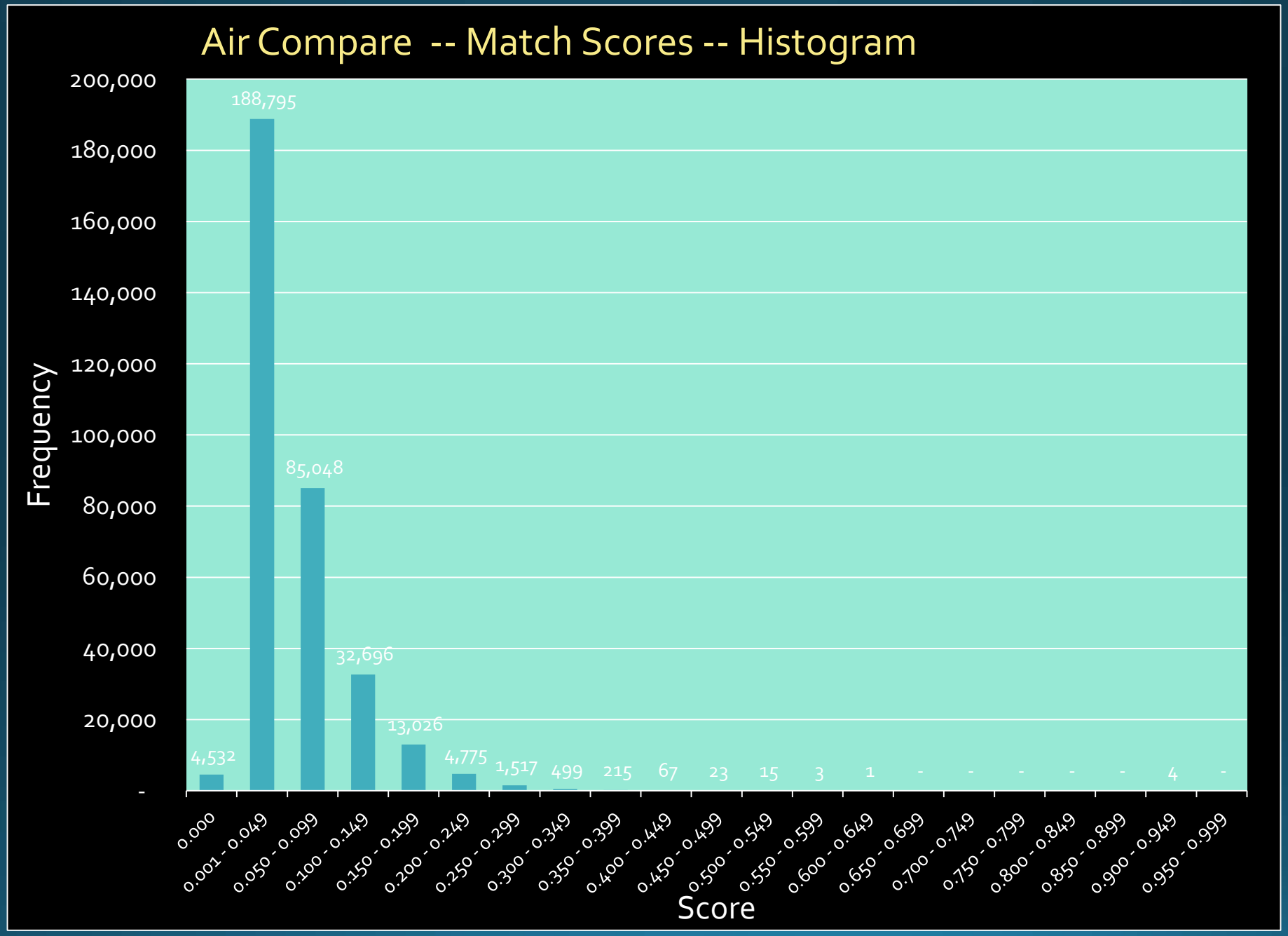
North Carolina Department of Environmental Quality

Goal of the Project

- Identify common data elements across the CAER systems
  - Determine whether it is possible to reduce duplicative reporting by industry
- Identify common value lists (aka code set lists or permissible value lists)

Process of Work

- Gathered data dictionaries for four national CAER systems and three state systems (IA, NC, TX)
  - Ensured all data elements had clear definitions
    - Referenced previously reviewed data dictionaries
  - Standardized formatting of data for comparison
  - Compared data elements using open source tool
  - Compared permissible values and permissible value lists
  - Worked with system owners to improve definitions of data elements and code set values
  - Prepared data dictionaries ultimately for loading into Data Element Registry Services (DERS)
    - Catalog of data dictionaries and code sets to promote reuse


Air Compare -- Match Scores -- Histogram
Frequency
200,000
180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
-
4,532
188,795
85,048
32,696
13,026
4,775
1,517
499
215
67
23
15
3
1
-
-
-
-
-
4
-
0.000
0.001 - 0.049
0.050 - 0.099
0.100 - 0.149
0.150 - 0.199
0.200 - 0.249
0.250 - 0.299
0.300 - 0.349
0.350 - 0.399
0.400 - 0.449
0.450 - 0.499
0.500 - 0.549
0.550 - 0.599
0.600 - 0.649
0.650 - 0.699
0.700 - 0.749
0.750 - 0.799
0.800 - 0.849
0.850 - 0.899
0.900 - 0.949
0.950 - 0.999
Score

Post Comparison Work

- Created Excel spreadsheet showing data elements that map or might map
  - Created subject-specific workbooks
    - Facility information; Substance information; Contact information; Address information
      - Aligned with Exchange Network data standards
    - "Other" workbook showed potential mappings for elements that did not fit the above categories
      - E.g., "Average annual days per week"
- Created Excel spreadsheet showing value lists that map or might map
- Worked with system contacts to validate findings
  - Removed false positive matches

Findings – Data Elements

- Great degree of commonality for:
  - Facility, Contact, Substance, Address, Location, Industrial classification
- Facility collected differently by NC
  - Facility module used by multiple NC systems
- Address collected differently across systems
  - Address sometimes its own table, used for both Facility and Contact
  - Address sometimes part of Facility table and part of Contact table
- Substance collected differently across systems
  - GHG uses the term "Gas", causing the comparison to miss this data element
  - TRIPS collects chemical names, synonyms and CAS numbers
  - NC collects CAS numbers and pollutant codes
  - EIS and Iowa collect chemical names and pollutant codes
  - CEDRI collects only pollutant names

Findings – Data Elements (cont'd)

- Very little commonality for "Other" data elements
  - CEDRI had the fewest common data elements with other systems
    - However emission release point data was found to be common/similar with EIS
  - TRI had few common data elements with other systems
    - Reporting year, Comment text
  - GHG program had some common data elements with other systems
    - However it had more overlap with EIS and NC than the other systems, especially for Emission Unit and Measurement elements
- Great degree of commonality between EIS and three state systems
  - Iowa had 181 common data elements with EIS, and 29 with other EPA systems
  - NC had 112 common data elements with EIS, and 61 with other EPA systems
  - Texas had 22 common data elements with EIS, and 17 with other EPA systems

Findings – Examples of Data Elements

- Facility Name
  - Name collected by all systems, except STARS
  - Texas' STARS collects a Site_ID
- Calculation method
  - EIS collects Emissions calculation method
  - GHG collects multiple calculation methods that are specific to the pollutant (ex: N20 Emissions Calculation Method) and the industry type (ie. Nitric Acid Production)

Findings – Value Lists

- 656 value lists were evaluated for potential matches, 71 lists were identified as having possible overlap and had their values compared
- NAICS Codes are collected across most systems
- County and Country codes could be standardized
- For CEDRI, there is overlap with only NC's Pollutant permissible value list
- For TRIPS, there is limited overlap with permissible values of other systems
- For GHG and EIS, there appear to be some permissible values that could be standardized
- The state value lists have significant overlap with EIS (15 value lists) and limited overlap with GHG and TRIPS

Findings – Examples of Value List

- NAICS codes are collected by most systems
  - Values of the lists differ, with each system uses slightly different subsets
- EIS Release Point Type and NC EMIS_REL_POINT_TYPE_CODE
  - Only 2 value lists to exactly match for both their codes and descriptions.
- GHG has overlap with NC for Control System Type and Unit Type
- EIS CountryFipsCode and TRIPS V_Country_FIPS were expected to have overlap.
  - However, EIS uses a combination of 2-letter Country codes and 2-digit state codes for Canada and Mexico, while TRIPS only uses the 2-letter Country codes in this list.

Next Steps

- Fold in ICIS-Air data dictionary into comparison
- Discuss with system owners how to adopt data shared services reduce reporting burden and promote integration
  - Facility Registry Services (FRS)
  - Substance Registry Services (SRS)
  - North American Industrial Classification System (NAICS) web services
  - Sharing other code sets as appropriate
- Work with system owners to determine if it is possible to standardize data elements and permissible values that seem to be the same

Questions & Comments

Contact

John Harman (harman.john@epa.gov)

Tammy Manning (tammy.manning@ncdenr.gov)

Lauren Gordon (gordon.lauren@epa.gov)

Thank you!