Employer Data Matching Workgroup White Paper

January 2017

CONTENTS

EXECUTIVE SUMMARY[1]

Matching and reusing data on employers across Federal government agencies can have multiple and significant benefits, but it is currently very difficult to do. In order to begin to address these issues, OMB convened an interagency Employer Data Matching Workgroup. The Workgroup was co-chaired by senior members of the Census Bureau and the Bureau of Labor Statistics, and included over 40 expert Federal staff across 14 Agencies, representing 29 program, evaluation, or statistical offices. The scope of this project included examining current and potential future methods of matching data at multiple levels (establishment, and firm or enterprise), matching parent/child relationships within firms or enterprises, and capturing the dynamic nature of these relationships as they change over time. Key tasks included:

- Documenting issues Federal agencies face related to matching and uniquely identifying establishments and firms within and between data sets and over time;
- Identifying current successful strategies and/or policies used by Federal agencies to address matching challenges in the context of analyzing data, conducting evaluations, producing statistics, and identifying where additional strategies may be needed to further facilitate this work; and
- Summarizing potential future steps Federal agencies can take to improve the Federal government's ability to identify and match unique firms and establishments (and the relationship between the two) within and across Federal data sets, for the purposes of analyzing data, conducting evaluations, and producing statistics.

Matching Employer Data

Many Federal administrative and statistical activities require a matching process. In general, matching activities fall into one of the following fundamental types of activities:

- **Finding data on the same entity within a single data set:** agencies are de-duplicating and aggregating data, within the same business level (for example, at the establishment or enterprise level) and within the same data set, to find all observations related to a single legal entity.
- **Aggregating data within a corporate structure**: agencies aggregate data to the enterprise level of a corporate structure in order to group all observations related to a single enterprise.
- **Matching microdata at the same business level** between two or more data sets for:
  - Statistical purposes (including program evaluation): For example, agencies add variables to existing data sets to enhance quantitative analyses of firm behavior.
  - Programmatic purposes: For example, agencies may use linked data sets to support decision making from merged data that better defines market activity, and resulting risks.
- **Matching between different business levels** in two or more data sets: for example, agencies link different business levels to more fully understand corporate structures in the context of successorship, franchising, and multisite employers, at a given point or over a period of time.

Types of matches

There are two types of matching: deterministic and probabilistic:

- **Deterministic**, or exact, matching, looks for an exact match between two pieces of data.

[1] This white paper is intended to provide the Commission on Evidence-Based Policymaking with background information on topics relevant to the Commission's work. The paper was prepared by staff from OMB and staff from other Federal agencies.

- **Probabilistic**, or "fuzzy," matching uses a statistical approach to assess the probability that two records represent the same entity.

Data quality is a key factor in determining which method to use for matching. If data are well-curated, deterministic matching is the simpler, more accurate, and faster method, when the two data sets contain the same unique identifiers to perform the match. Often, such identifiers are not available, or the identifiers present within the data sources do not uniquely identify the entities to be matched. In such cases, deterministic matching may still be possible, but only with painstaking research for each case. Probabilistic matching is more complex than deterministic matching, but it provides an approach for matching when deterministic matching is not feasible. It is often difficult and resource-intensive to evaluate the quality of probabilistic matches.

Challenges in matching employer data

There are two primary issues that drive the vast majority of the challenges in matching data: the lack of a common universal identifier for employer units and poor quality of the underlying identifying data.

The greatest barrier to matching data on employers across data sets is the lack of a common, or universal, business identifier. Eliminating this obstacle by developing a Federal system to create and manage a universal identifier could result in cost savings in matching but would require a major investment of time and Federal resources to create and maintain such an infrastructure. Assuming that the identifier could be created, it would be a challenge to enforce consistent use of such an identifier by all employers on the domestic and international fronts. This identifier would need to capture various corporate/industry levels and change over time (in other words, it should change with firm births, deaths, mergers, acquisitions, etc.), and no Federal entity has the authority, staff, or resources to collect and manage such information.

There are examples of voluntary, widespread adoption of important taxonomies, such as the North American Industry Classification System (NAICS). Given that the creation and use of a universal identifier is likely in the best interest of taxpayers and will likely reduce Federal and enterprise burden, it would be worth exploring whether a voluntary means of adoption is viable. Such voluntary adoption may also be complicated by the nature of the global economy and the domestic and international structure of some employers.

Because there is no universal employer identifier which meets cross-agency needs, agencies often have to expend significant resources to research each case for deterministic matching or to obtain data for probabilistic matching. That is, agencies can and do combine different data fields to match employer data, but the effectiveness of this approach varies based on data fields available and data quality in those fields. Common issues include: missing important data fields, inconsistent data formats, and change over time in critical matching fields.

Best Practices

There is substantial potential to achieve efficiencies in matching U.S. employer data across Federal data sets for data analysis, evaluations, and statistical activities, based on common needs across agencies. To this end, there are immediate steps agencies can take to adopt best practices for data collection and matching, which are illustrative of the nature of practices Federal agencies have developed to deal with difficult, entrenched matching challenges. There are also ways to maximize the use of existing authorities and data sets to enable matching. In particular, leveraging data elements common among Federal agencies which, in combination, constitute a universal unique identifier, would facilitate efficiencies for matching Federal data sources.

Notably, the utility of Federal data sources can be increased by including as many cross-agency identifiers as possible. While these fields are useful for matching, it is important to note that identifiers have confidentiality, privacy, and proprietary concerns.

Moving forward, there are topics that could benefit from further investigation from the Employer Data Matching Workgroup, such as the development of a roadmap for implementing common data elements in Federal data sources, the development of an authoritative source of business existence and characteristics for Federal agencies, the possible role and value of a centralized data sharing "referee," and the establishment of a Federal community of practice for matching and entity resolution.

INTRODUCTION

The Federal Government currently collects data from employers and enterprises in the United States for a wide range of purposes, including administering small business loan programs, administering regulatory requirements, and producing valuable economic statistics. While these data collections are valuable and frequently necessary, in some cases they can result in the collection of duplicative information from employers. For example, both the Bureau of Labor Statistics and Census Bureau collect duplicative information on businesses with multiple locations.[2]

Several agencies have taken steps to attempt to reduce this burden while also increasing the amount and usefulness of the information generated from these data collections.[3] For example, since the 1990's, agencies within the Federal Statistical System have taken advantage of electronic data and web-based reporting to minimize respondent burden and significantly lower data processing costs.[4] Additionally, the Environmental Protection Agency (EPA) implemented an Application Programming Interface (API) that allows regulated establishments and facilities to take advantage of previously-reported information for querying, retrieving and pre-populating future required reports, which not only improved the data quality and the ability to match data across multiple programs, but also reduced the annual reporting burden by over 140,000 hours for a single EPA program.[5]

---

[2] The BLS conducts its quarterly Multiple Worksite Report (MWR), which asks most multi-location employers to provide monthly employment and quarterly wage data for all of their establishments covered under one Unemployment Insurance (UI) account in a state. Most multi-location employers with a total of 10 or more employees combined in their secondary locations are required or requested to complete the MWR. The Multiple Worksite Report is designed to collect information showing the distribution of the monthly employment and quarterly wages of business establishments, by industry and geographic area. Information on the MWR form is used to more accurately classify employment and wage data of multiple establishment employers by industry and by location within a State. By collecting and storing employment and wage data by worksite, states and the BLS can disaggregate these data below the county level for more extensive and detailed analysis of business and economic conditions within their state, including local and regional employment totals. These data are used to ensure an equitable distribution of federal funds through grant programs that use county economic indicators as a basis for allocations. No other sources are available to obtain this information.

The Census Bureau conducts its annual Company Organization Survey (COS) to obtain similar information on multi-establishment firms in order to maintain its Business Register (BR). Annual data collection for the COS begins in late December of the reference year for the pay period of March 12th. Reported data are for activities taking place during the reference year. An annual mail-out survey of selected companies is conducted for large multi-establishment companies with 500 or more employees are selected with certainty. Small multi-establishment and single-establishment companies are selected based on administrative data indicating a probable organizational change. All selected companies are identified from those maintained on the Business Register. Survey results are available to the Census Bureau about 8 months after each reference year and are used throughout Census Bureau economic data program operations, as a major source of information for County Business Patterns reports, and as a resource in responding to requests for a variety of special reports and reimbursable tabulations.

Recently, BLS and Census have been sharing their multi-unit data which accrues cost efficiencies and improves the comparability and accuracy of Federal economic statistics by improving the consistency of multi-location data and reducing respondent burden.

[3] The examples noted represent the current range of burden reduction activities occurring among Federal agencies. By achieving efficiencies in matching through the white paper's suggested approaches, future burden reduction is possible (e.g. avoiding new surveys).

[4] For example, see:
https://www.census.gov/history/www/innovations/data_collection/counting_the_population.html
http://stats.bls.gov/opub/mlr/2016/article/one-hundred-years-of-current-employment-statistics-data-collection.htm

[5] EPA Toxic Release Inventory TRI-MEWeb 2.0 Reporting Burden Estimate

At the same time, there is a growing interest in identifying more efficient and effective ways to help employers succeed and comply with Federal requirements. One cost-effective method to generate these insights is by analyzing and evaluating the effectiveness of existing programs using the data that government already collects. Frequently the programs that are being evaluated are designed to help firms grow, ease compliance with Federal requirements, or promote innovation. However, the data on the outcomes of interest, whether it is employment growth, regulatory compliance, economic growth, or competitiveness, frequently reside in different government data sets, across multiple agencies. To capture this outcome information, agencies can either collect the information again—at additional cost and burden on employers, or they can get the data from a data set elsewhere in government.

Similarly, statistical agencies may be able to more cost-effectively and efficiently collect and generate economic statistics by re-using already-collected data. Using such data, statistical agencies can establish sampling frames which are more cost-effective by accounting for additional information on the sampling units. Additionally, improvements in administrative data can reduce burden associated with statistical collection by reducing the need for duplicate reporting of information. If statistical agencies can gain access to key data via the administrative data that employers already report as a regular part of their business activities, there will be less need for the statistical agencies to ask employers for those data again through surveys. Further, combining these data can lead to new, improved, and valuable data products and evidence that can benefit employers and help government develop smarter policies[6].

It is currently very difficult for agencies to access and use this information for statistical, evaluation, or other purposes[7]. In order for these data to be useful, researchers, statisticians, and evaluators must be able to match data on individual employers within the same data set and between different data sets, and they must be able to do so reliably and with little error. Access to the data may also be limited by law, such as by Title 26 or the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA). This type of matching requires the ability to accurately identify the same employer in two or more different data sets. When conducting this type of match for individuals, frequently a combination of unique identifiers is used, including the social security number (SSN), name, and date of birth (DOB). However, for employers, there is no standard equivalent to SSN, name, and DOB. In some cases this variation in employer identification approach is the byproduct of different program rules and requirements. In other cases it is an artifact of the way government data systems have developed over time. In all cases, this variation is a large barrier to determining if Federal programs and policies affecting employers and their employees are effective, especially when answering such questions requires data from multiple data sets. Just a few of the complicating factors involved include:

- Lack of a unique identifier. There is no government-wide policy on which, or even whether, any unique identifier should be used across all government data sets that affect employers. Data sets may include different identifiers for legal, policy, or programmatic reasons, such as to maintain consistency with historical data or to minimize burden on respondents.
- Inconsistent level of identification. Businesses have multiple levels of organization, including the establishment (physical location), enterprise (which may consist of multiple establishments), and parent company (which may consist of multiple enterprises). There may also be differing needs for ultimate domestic corporate parent versus global corporate parent. Depending on the purpose of a program or statistical collection, Federal agencies track employer information at different levels of identification, and matching across these levels can be difficult.

[6] See also: "Using Administrative and Survey Data to Build Evidence." July15, 2016. https://www.whitehouse.gov/sites/default/files/omb/mgmt-gpra/using_administrative_and_survey_data_to_build_evidence_0.pdf
[7] See also: "Barriers to Using Administrative Data for Evidence-Building." July 15, 2016. https://www.whitehouse.gov/sites/default/files/omb/mgmt-gpra/barriers_to_using_administrative_data_for_evidence_building.pdf

- Data quality. Data quality issues, such as respondent reporting error, or lack of formatting consistency among fields for matching, prevents or complicates the matching processes.
- Timing. The timing for when the data are collected may complicate the matching process as more current information may be available for the same unit of observation.

For all of these reasons, it is often highly labor intensive and expensive to match and reuse these data. Matching across data sets can take months or even years, and still may not fully serve the intended goals. These problems occur across statistical, evaluation, and program functions in Federal agencies. While the purpose of matching may differ, the matching and analytical challenges are common across these functions.

Overview of the Workgroup

In order to begin to address these issues, OMB convened an interagency Employer Data Matching Workgroup. The Workgroup was co-chaired by senior members of the Census Bureau and the Bureau of Labor Statistics, and included over 40 expert Federal staff across 14 Agencies, representing 29 program, evaluation, and statistical offices. The scope of this project included matching at multiple levels (establishment, and firm or enterprise), matching parent/child relationships within firms or enterprises, and capturing the dynamic nature of these relationships as they change over time.[8]

The Workgroup was charged with developing strategies to make it easier to match data on U.S. employers across Federal data sets for statistical purposes. **Statistical purposes** refer to the use of data to better understand the characteristics, behavior, or needs of groups.[9] Program evaluation falls within this definition. Statistical purposes exclude uses of data that affect the rights, benefits, or privileges of individual entities: indeed one of the defining characteristics of statistical uses is that data about an individual entity are never made public and are never used to make decisions about that entity. But statistical purposes include a wide range of analytic uses, where only aggregated and de-identified data are made public.

While the Workgroup focused its activities on these statistical purposes, it became clear that many of the best practices identified could have benefits for non-statistical purposes. One of the primary ways this work could reduce burden on employers is by facilitating the re-use of individual-level employer identifying data from one program to another (e.g., by allowing an employer to select itself from a pre-populated list or auto-loaded list of establishments or entities rather than entering the data anew each time). This type of activity requires identification and release of information at the individual identity level, so even though the information may be otherwise publicly available (e.g., name and street address), it falls outside of the definition of a statistical purpose.

Workgroup members have spent many years finding ways to improve matching for their specific purposes. The goal of this white paper is to share knowledge and best practices the Workgroup believes would yield potentially significant benefits in reducing agency workloads, burden associated with statistical collection, and reporting burden for employers.

---

[8] See Appendix D for a detailed description of participating Federal agencies and the Workgroup methodology.

[9] Note that a statutory definition of "statistical purpose" exists in section 502(9) of CIPSEA:
"The term 'statistical purpose'-
"(A) means the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups; and
"(B) includes the development, implementation, or maintenance of methods, technical or administrative procedures, or information resources that support the purposes described in subparagraph (A)."

WHAT IS AN "EMPLOYER?"

In order to identify the barriers and potential best practices for matching and reusing data on employers, it is first essential to identify what we mean by "employer." This task is somewhat more complicated for businesses than for individuals, given that the identification of a person is fairly stable regardless of context. However, there are multiple levels of business that may be of interest. For example, one could be interested in the effect of a technical assistance grant on a single business location, or the effect of a small business loan on a firm that has multiple locations.

Terms such as "employer," "firm," or "establishment" are often defined differently in statutes and regulations across Federal agencies and programs. Many of these terms are based on commonly shared concepts, so while precise matching of terms across agencies is not possible, some generalization is both useful and practical. For clarity, this white paper uses an internal taxonomy to account for the definitional differences among Federal agencies:

- **Establishment:** a single physical workplace or facility. An establishment is commonly understood as a single economic unit, such as a farm, a mine, a factory, or a store, that produces goods or services, for which payroll and employment records are kept. Establishments are typically at one physical location and engaged in one, or predominantly one, type of economic activity for which a single industrial classification may be applied.[10]
- **Legal Entity:** a legal person or structure that is organized under the laws of any jurisdiction.[11]
- **Enterprise:** Alternative terms for enterprise are *firm* and *company*. An enterprise is a legal business entity that may consist of one or more establishments. Each establishment may or may not participate in a different economic activity. The establishments may have different physical addresses. If they do, the physical address of record should be for the headquarters or main office of the enterprise.[12]
- **Parent Company:** An enterprise that owns all or the majority (51% +) of another enterprise so that the latter stands in relation to the former as a subsidiary.[13]
- **Employer**: a legal entity or individual identified as a worker's employer either nominally (for example, on the workers' paychecks or tax forms) or through an employment relationship. This term can pertain to establishments, enterprises, and parent companies.[14]

Federal data sources and related matching efforts often capture relationships among establishments, enterprises, and parent companies, for mission-related analysis and reporting. For example, an establishment-based enforcement program might collect all of their data at the establishment level, but in order to analyze trends by enterprise or industry, need to be able to aggregate those data. As another example, a statistical agency may collect data at the establishment level, but publish tabulations of these data by the size of the parent company.

---

[10] Portions of this definition are derived from the definition of "establishment" within the Bureau of Labor Statistics' Quarterly Census of Employment and Wages (QCEW) Business Register (BR), and the Census Bureau's Business Register.

[11] This definition is derived from the International Organization for Standardization's definition of Legal Entity Identifier. For further information, see: http://www.iso.org/iso/catalogue_detail?csnumber=59771

[12] Portions of this definition are derived from the definition of "firm" or "company" within the Bureau of Labor Statistics' Quarterly Census of Employment and Wages (QCEW) Business Register (BR), and the Census Bureau's Business Register's definition of "enterprise."

[13] From: https://www.dol.gov/vets/contractor/main.htm#20

[14] The definition of "employer" under the Fair Labor Standards Act informs this definition.
See also: https://www.dol.gov/whd/regs/compliance/whdfs13.pdf

MATCHING EMPLOYER DATA

Why is matching conducted?

Many Federal administrative and statistical activities require a matching process. For example, if an agency wants to avoid having duplicate observations for an establishment within an administrative database, then it has to identify all reports in its data that apply to a specific establishment, and then look to see if there are duplicates for certain phenomena. In general, matching activities fall into one of the following fundamental types of activities:

- **Finding data on the same entity within a single data set:** agencies are de-duplicating and aggregating data, within the same business level (for example, at the establishment or enterprise level) and within the same data set, to find all observations related to a single legal entity. For example, agencies regularly undertake projects, involving searches within administrative data sources, to identify all instances of transactions with particular legal entities. A regulatory agency may search its administrative data to determine the compliance history of a particular employer.
- **Aggregating data within a corporate structure**: agencies aggregate data to the enterprise level of a corporate structure in order to group all observations related to a single enterprise. For example, agencies will nest companies within larger, related aggregates in order to measure economic activity over time. Statistical agencies may aggregate data (e.g. employment counts) for branch locations of an enterprise, in order to measure the size of the enterprise.
- **Linking/Matching microdata at the same business level** between two or more data sets for:
  - Statistical purposes (including program evaluation): For example, agencies add variables to existing data sets to enhance quantitative analyses of firm behavior. A statistical agency may integrate data sources to augment one source with fields not initially contained within the source, such as firm age.[15]
  - Programmatic purposes: For example, agencies may merge data sets to provide fuller information on market activity and risks to support decision making. A regulatory agency may take advantage of merged employer data sources to get a more complete picture of firms and sources of financial risks in an industry.
- **Linking/matching between different business levels** in two or more data sets: for example, agencies link different business levels to more fully understand corporate structures in the context of successorship, franchising, and multisite employers, at a given point or over a period of time. An agency may match data on corporate hierarchies with another employer data set, to further understand the employer's relationships with other legal entities. Also, an agency may tabulate data collected at the establishment level by characteristics of parent companies.

[15] Additionally, statistical agencies link microdata at the same business level to develop new data products. For example, BLS's QCEW program matched publicly available IRS data on nonprofits with its business register to develop new data on the non-profit sector. This initiative is meeting the needs of data users with no new resources while imposing no new respondent burden on businesses. These new combined research data covering 2007–2012 were released in September 2014, meeting a longstanding need for recent, detailed industry and geographic detail on this large sector of the economy. The nonprofit sector covers about 10% of employment and has higher than average wages, making this segment of the economy important to understand. Another BLS matching project overlays hurricane flood zones over geocoded business locations. Maps and tables are published and available on the BLS website showing the number of establishments, and the accompanying employment and wages that are exposed to potential damage under hurricane conditions of varying strengths.

Types of matches

There are two types of matching: deterministic and probabilistic:

- **Deterministic**, or exact, matching, looks for an exact match between two pieces of data. In order for this method to be effective, the data being matched should uniquely identify the entity of interest, and the same data field should be present and formatted in the same way in both records. This method is ideal if the data sets of interest are reliably and accurately collecting unique identifiers, and the unique identifiers are well-matched to the purpose of the analysis.
- **Probabilistic** matching uses a statistical approach to assess the probability that two records represent the same entity. In order to accomplish this, a set of data fields are compared between two records and the closeness of the match between two record pairs is assessed.

Data quality is a key factor in determining which method to use for matching. If data are well-curated, deterministic matching is the simpler, more accurate, and faster method, when the two data sets contain the same unique identifiers to perform the match.[16] Often, such identifiers are not available, or the identifiers present within the data sources do not uniquely identify the entities to be matched. In such cases, deterministic matching may still be possible, but only with painstaking research for each case. Probabilistic matching is more complex than deterministic matching, but it provides an approach for matching when deterministic matching is not feasible. It is often difficult and resource-intensive to evaluate the quality of probabilistic matches.

Deterministic and probabilistic matching methods have the potential to produce data structures that give additional insights the original data could not have provided. Several agencies and offices reported the use of matching methods to turn administrative data sources, with multiple discrete observations of employers or establishments, into quasi-longitudinal or time series data, by linking observations over time. Similarly, there are matching applications that can link subsidiaries to parents or nest companies within larger, related aggregates.

Current unique identifiers used across Federal Agencies

Several entities have created unique identifiers (IDs) in order to improve matching and identification of employers, which are used across multiple Federal Agencies. However, none of these IDs are universally collected, and none of them uniformly identify the level of business or the relationship between the levels of business. There are four primary unique identifiers currently in use by Federal Agencies: the Employer Identification Number (EIN), Data Universal Numbering System (DUNS©) numbers, Commercial and Government Entity (CAGE) codes, and the Legal Entity Identifier (LEI). To apply for a Federal contract or grant, an entity must have an EIN, DUNS© number, and CAGE code. However, these identifiers are not currently used throughout all Federal data sets as only a small percentage of U.S. enterprises register annually for Federal contracts or grants. These identifiers enable clear identification of unique entities seeking Federal dollars and are used to identify exclusions, past performance history, and business integrity. Additionally, the LEI, as required under Public Law 111-203 (commonly known as the Dodd-Frank Wall Street Reform and Consumer Protection Act), is heavily used within the federal financial regulatory community, and by non-financial regulatory agencies.[17]

[16] See also: https://www.healthit.gov/archive/archive_files/FACA%20Hearings/2010/2010-12-09%20Patient%20Linking/Probabilistic%20Versus%20Deterministic%20Data%20Matching.pdf

[17] For a complete list of Federal agencies currently using LEI, see: https://www.gleif.org/en/about-lei/regulatory-use-of-the-lei.

- The **Employer Identification Number (EIN)** is issued by the IRS. Under the Internal Revenue Code (IRC), every US entity is required to have an EIN for tax purposes regardless of whether they have Federal contracts or grants. An enterprise subject to Federal income tax will file using this EIN on its own separate tax return or file under the Parent Company's EIN of a consolidated return if it elects to file with an affiliated group of other enterprises.[18] Furthermore, establishments can be associated with multiple EINs or an enterprise could use the same EIN for all its establishments.
- A **Data Universal Numbering System (DUNS©)** number is a uniform and unique nine-digit number administered by Dun and Bradstreet (D&B). The number is assigned by D&B and is currently collected by the Federal government as part of the registration process for grants and contracts.
- A **Commercial and Government Entity (CAGE)** code is a uniform and unique five-character alpha-numeric identifier for entities in the U.S. CAGE codes are used internationally as part of the North Atlantic Treaty Organization Codification System (NCS). Management of CAGE codes in the United States is done by the Department of Defense. If an employer has not applied for Federal contracts or grants, the entity would not need a CAGE code and thus may not have one.
- The **Legal Entity Identifier (LEI)** is a 20-digit, alpha-numeric code based on the ISO 17442 standard developed by the International Organization for Standardization (ISO). It connects to key reference information that enables clear and unique identification of legal entities participating in financial transactions, such as those participating in financial transactions or when used in regulatory and supervisory reporting.[19]

There are a number of other intra-Agency identifiers which Federal Agencies use for internal databases, or for limited cross-agency coordination. For example, the **State Unemployment Insurance Account Number** (UI number) is issued by state unemployment insurance agencies. The state identification number is assigned by each state to identify employers covered by State UI laws or to identify federal government installations covered by Unemployment Compensation for Federal Employees (UCFE) provisions. UI account numbers are utilized at the federal level to identify establishments maintained in the BLS's Quarterly Census of Employment and Wages (QCEW) files. This field is consistent from quarter to quarter and allows for identification of the same unit over time.

Challenges in matching employer data

There are two primary issues that drive the vast majority of the challenges in matching data: the lack of a common universal identifier for employer units, and poor quality of the underlying identifying data.

Universal Employer Identifiers

Agencies lack a single, universal identification system for establishments, firms and other types of employer units in Federal and non-Federal data. In the ideal system, such an identifier would be hierarchical such that each employer unit, at each level (i.e., establishment, enterprise, parent company) would have its own unique ID, and the set of identifiers would be used together to identify the relationships among the levels over time.

[18] Further, some foreign entities that are US-owned have an EIN for Federal tax purposes.
[19] "Introducing the Legal Entity Identifier (LEI)." Global Legal Entity Identifier Foundation. <https://www.gleif.org/en/about-lei/introducing-the-legal-entity-identifier-lei>. Accessed 12/9/2016. See also: http://www.leiroc.org/lei.htm.

Thus a single establishment could be tracked not only across multiple data sets, but also as it is sold from one enterprise to another.[20]

If such an infrastructure existed, linking would be simple, as deterministic matching with a single data element would be feasible, and connections among different business levels over time would be implicit in the identification system. Developing such an infrastructure could result in cost savings over time for both statistical and program agencies, though more work needs to be done to compare the costs and benefits of developing this infrastructure.[21] Such a system would also enable an employer to provide a piece of data to government only once, rather than multiple times, subject to any legal limitations on sharing of the data. A primary challenge with such a system is maintaining the accuracy and currency of the underlying firm or employer data. While matching would be simple, verifying the mergers, acquisitions, incorporations, and continuous changes would require significant domestic and international resources or connections. Absent such due diligence, employers could change incorporation or legal structure, and obtain new identifiers. Assigning improper activity, sanctions, or other such determinations based on inaccurate identification could have serious repercussions.

There is great variability among statutes, regulations, agency policies, and reporting definitions in how employer units are identified. This variation largely comes from differences in laws and policies as they apply to different subsectors of business. For example, the Fair Labor Standards Act of 1938 (FLSA), which "establishes minimum wage, overtime pay, recordkeeping, and child labor standards affecting full-time and part-time workers in the private sector and in Federal, State, and local governments,"[22] defines "employer" to include "any person acting directly or indirectly in the interest of an employer in relation to an employee and includes a public agency, but does not include any labor organization (other than when acting as an employer) or anyone acting in the capacity of officer or agent of such labor organization."[23] In comparison, the Occupational Safety and Health Act of 1970 (OSH Act), which "assure[s] safe and healthful working conditions for working men and women," defines "employer" to mean "a person engaged in a business affecting commerce who has employees, but does not include the United States (not including the United States Postal Service) or any State or political subdivision of a State."[24]

The statutes have additional differences, such as differences in coverage. A 2002 GAO study noted when comparing the FLSA and OSH Act that: "Coverage under the OSH Act is broader. All employees of a particular employer are covered if the employer is engaged in a business affecting commerce. Coverage under the OSH Act does not depend on the specific activities of the employee or the volume of the employer's business."[25]

This variability has obvious ramifications for what can reasonably be expected from matching data sources from very different programs and approaches to defining "employment," "employer," and "employee," as well as corporate structures such as "firm," "establishment," and "enterprise." [26] Many of these inconsistencies would still present obstacles to matching even if the government were to develop a common conceptual standard and definition for the unique identifiers.

---

[20] More specifically, databases with such a universal unique identifier would ideally contain a history for each establishment of its relationships over time (i.e. its history of hierarchical identifiers). That database would allow for tracking of the establishment over time even if it changed owners and relationships.

[21] Costs could include, for example, staff to research the incorporation or legality of an entity, build corporate family trees, and track changes among corporations. Benefits could include elimination of costs related to poorly matched or unmatched data.

[22] See: https://www.dol.gov/whd/regs/compliance/hrg.htm

[23] See: https://www.dol.gov/whd/regs/statutes/FairLaborStandAct.pdf

[24] See also: https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=OSHACT&p_id=2743

[25] From: http://www.gao.gov/assets/240/235612.pdf

[26] The paper contains an internal taxonomy to ensure that there is a common set of definitions to support its description of best practices.

In addition, considering the variations in statutes, policies, and uses, creation of a new Federal infrastructure that would establish, maintain, and assign identifiers to commercial and Federal entities would require considerable effort. Such a change would require the establishment of a Federal program to manage the identifiers. In addition, such a change could require changes in statute to harmonize definitions of different employer units as well as modifications to policy and numerous data collection and maintenance systems across Federal government agencies. Such changes in program definitions would need to be handled in a manner so as not to disrupt the continuity of longitudinal data that is already being collected. Also, regulatory agencies could not leverage confidential data sources within Federal statistical agencies even if those sources contain a universal identifier.

Beyond cost and conceptual challenges, it would also be difficult to enforce consistent use of such an identifier by all employers within current authorities. There is no single Federal agency that would be a natural owner to enforce the use of such an identifier because no Federal entity currently collects all of this information. There are examples of voluntary, widespread adoption of important taxonomies, such as the North American Industry Classification System, or NAICS. However, this is a very broad standard that is applied to classes of enterprises and employers rather than used to uniquely and uniformly identify individual employers. Given that the creation and use of a universal identifier is in the interest of businesses and taxpayers, it would be worth exploring whether a voluntary means of adoption of a universal identifier is viable.[27]

Moreover, there are ways to maximize the use of existing authorities and data sets to enable matching. In effect, these probabilistic methods rely on a variety of existing, widely collected data fields, which when combined can effectively create a flexible "universal identifier" that can be adapted to different applications. These methods generally require the data to be of high quality and sufficiently standardized in order to be effective.

Because there is no universal employer identifier which meets cross-agency needs, agencies often have to expend significant resources to obtain data for probabilistic matching. Agencies currently deal with challenges related to:[28]

- **Resource and capacity constraints.** Acquiring data housed by other agencies can be costly and difficult for Federal agencies. The process of developing interagency agreements can be time-intensive, and agencies require employees with the proper skill set and time available to reconcile data with different definitions. Also, because many agencies do not have a complete and current data inventory, they may not be fully aware of data that is available to support their particular needs. After receiving the data, agencies may not know the usefulness of the available data until after significant efforts to review the source, depending on the availability of relevant documentation and metadata. Additionally, infrastructure may be a barrier to obtaining data. For example, if the construction of an agency's own or preferred data center is considered sufficient for one survey but not another (e.g., differences in requirements for physical security and data access protocols between two data sources might prohibit linkage), an agency may have to expend significant resources to determine an alternative strategy for accessing data. Lastly, the cost of purchasing external data sources can serve as a practical barrier for agencies with limited funding available for such purchases.

[27] A review of voluntary adoption would require the selection or scoping of a prospective identifier. The Workgroup puts forward a framework for Federal data sources to include existing, widely collected data fields, which when combined can effectively create a flexible "universal identifier" that can be adapted to different applications.

[28] See also: https://www.whitehouse.gov/sites/default/files/omb/mgmt-gpra/barriers_to_using_administrative_data_for_evidence_building.pdf

- **Legal barriers.** While there are Federal data sets containing information that can be valuable to regulatory agencies, limitations on the use of such data frequently restrain access. For example, under CIPSEA, data, including business data, acquired by an agency under a pledge of confidentiality and for exclusively statistical purposes can only be used by officers, employees, or agents of the agency exclusively for statistical purposes. Lacking access to an authoritative source on business existence and structure is a critical issue for correctly matching data, because agencies have no way to know if the data they are collecting is fully accurate. An authoritative source would be useful to serve as a "spine" to match data against.
- **Policy and legal interpretations.** At the agency level, there can be confusion in the interpretation of statutes such as CIPSEA as well as other policies on data sharing, which may create additional barriers for agencies trying to access and share Federal data sources for matching purposes. Agencies may spend months or years coming to agreement on the proper interpretation of a particular statute or policy, and to develop interagency agreements to allow matching. OMB's M-14-06, *Guidance for Providing and Using Administrative Data for Statistical Purposes*[29], encourages Federal departments and agencies to promote the use of administrative data for statistical purposes and provides guidance in addressing legal and policy requirements for such uses. It creates "a presumption in favor of openness to the extent permitted by law and subject to privacy, confidentiality, security, or other valid restrictions."

Data quality

Agencies can and do combine different data fields to match employer data, but the effectiveness of this approach varies based on data fields available and data quality in those fields. These issues can prevent deterministic matching for a number of reasons:

- **Missing important data fields:** Agencies frequently do not collect enough of the most important pieces of data to enable matching, and so they do not engage in data matching for those data sets. Federal agencies have generally built data collections for narrow and specific purposes, designing definitions, content, and formats to suit the immediate needs of the program performing the collection. While most sources with employer data include common data elements such as name, address, and basic characteristics such as NAICS code, these fields are not sufficient to match in all cases. For example, a match can be extremely difficult if not impossible to complete when data in the first data set is at the corporate level, but the matched data set is at the establishment level and does not include fields linking the establishments to their corporate hierarchies, such that neither data set has sufficient fields to create a cross-walk between the two. In some cases, agencies may have sufficient data to match on a deterministic or probabilistic basis, but may not have sufficient supplementary data to assess how good the match is.
- **Inconsistent data formats:** Federal agencies often have to deal with data quality issues arising from inconsistent data format standards. To cite a fairly common example, establishment names can vary considerably within and between data sources. Most data sources collect this information in free-text fields without edit checks, and allow variation. In extreme cases, establishment names can vary considerably, with a single Post Office branch being identified as "USPS", "US Postal Service", "United States Postal Service", and "USPS - MAIN STREET SOUTHBEND." In this case, deterministic matching of data sources by establishment name will not work, and even probabilistic matching methods may not work completely. Also, there is significant variation in data quality due to a lack of internal checks or naming conventions.

[29] https://www.whitehouse.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf

- **Change over time in critical matching fields:** With both deterministic matching and probabilistic matching, a common assumption is that the fields are static over time—that is the field is collecting the same information in year one as it is in year ten. However, there are a number of instances in which this is not the case, where fields are redefined over time. The business universe is dynamic, and issues such as relocation, mergers, successorship, and firm births/deaths produce a mismatch between data collection and changes in business status. For example, this issue arises in analyses of young, small businesses which may not have been in existence long enough to meet annual reporting requirements for collections. Evolving classification systems, such as NAICS and LEI, also present a unique challenge in matching data sources where observations occurred over a period of many years.

BEST PRACTICES

The Workgroup identified a number of practices that agencies could adopt within current authorities that have the potential to reduce burden on employers and agency staff while also facilitating better use of information. The Workgroup prioritized best practices based on the following criteria:

- **Efficiency or Cost-Effectiveness:** Examine the levels of effort and pecuniary costs of implementing each solution, and select the most efficient and cost-effective options.
- **Applicability:** Select the most useful methods, fields, standards, and data sets.
- **Relevance to Scope:** Use the methods, standards and data sets that are the most relevant to improving the matching process.
- **Alignment:** The Workgroup emphasizes best practices which best align to existing guidance and data standards, including guidance from, but not limited to, the National Information Exchange Model (NIEM), the International Organization for Standardization (ISO), and OMB memo M-13-13, "Open Data Policy – Managing Information as an Asset."

The Workgroup identified four practices that Federal programs and agencies may find useful in making it easier to match employer data:

Best Practices for Matching Data

The Workgroup identified a set of collection and algorithmic approaches that constitute the best practices in matching data on employers. Individuals conducting matches could use these methods rather than attempting to create a new method each time a match is to be conducted. The specific methods that produce the best results vary depending on the purpose of the match and the data fields that are available. These methods are described in greater detail in Appendices B and E, and a related bibliography is available in Appendix C.

It is worthwhile to note that the quality of a match depends largely on the quality of the underlying data in each data set. Before conducting a match, individuals should engage in data cleaning and standardization. In addition, employing a range of matching algorithms is critical to obtaining optimal matching results. Finally, individuals conducting a match should take advantage of data quality analysis; improvements in business rules, which improve data quality (by permitting data to be entered into a system if they meet certain criteria); and improved use and creation of data documentation to improve matching outcomes.

Data Inventory

One of the common challenges to matching employer data is the identification of data sets that are available for matching and what those data sets contain. To ease this identification process, the Workgroup identified a representative sample of the most frequently desired or most valuable data sets that the government currently collects. These data sets contain information on individual employers, firms, and/or establishments that have the widest coverage, greatest use, or which agencies are most interested in matching to other data sets. This data inventory is included as an attachment to the white paper. The data inventory includes information on the data sources' coverage of U.S. businesses, collection methodology, access restrictions, and information related to common fields for matching. CIPSEA limits the use of many of these data sources exclusively to statistical purposes.

Common Data Fields

Federal agencies can benefit from adopting the common data fields for matching purposes detailed in Table 1.[30] These fields are most commonly used in the methods described above and in Appendix B. Table 1 also accounts for existing guidance from other data-sharing initiatives (such as for Federal spending transparency data standards, and the National Information Exchange Model, or NIEM), and lessons learned from prior studies linking administrative and statistical data sources.[31] [32] [33] These fields can provide an exact match, and aid in matching even if an exact match is not possible due to variance amongst data sources.

Table 1 proposes a set of "Tier 1" fields, which are the most essential fields for matching and entity resolution. Agency data sources lacking these fields are very difficult to match, and the Workgroup believes Federal Agencies can achieve potential cost savings if their data sources universally incorporated Tier 1 fields. It is possible to compensate for missing fields by creating a matching profile from a combination of Tier 1 and non-Tier 1 data elements. Please note the following considerations for the Tier 1 fields:

- **Establishment-Focused:** Employer data is collected at various organization levels. The proposed framework emphasizes establishment names and addresses to enable matching with potential authoritative sources described in Appendix A. However, many Federal employer data sources do not always use establishment as the unit of analysis. Placeholders for other fields to describe organization level are included as Tier 2 fields.
- **Treatment of Identifiers:** Generally, the utility of Federal data sources in regards to matching, linkage, and reuse can be increased by including as many common identifiers as possible. Table 1 therefore includes a number of data elements to reflect this prioritization. Additionally, while these fields are useful for matching, it is important to note that identifiers have confidentiality, privacy, and proprietary concerns. Where possible, Federal agencies should include non-sensitive and non-proprietary identifiers. Ideally, Federal agencies would institute policies and processes that support consistent identification of employers within an agency. This would minimize the data cleansing necessary to identify the same employer involved in programs across Federal agencies.

These fields are generally common among 39 data sources in the data inventory. Federal data sources within this categorization commonly have multiple Tier 1 data fields available for matching, including: at least one identifier, legal and trade names, establishment physical location and mailing addresses, country codes, and data timestamps. Table 1 also shows web sites and e-mail addresses as Tier 1 fields.

Additional helpful fields include: information on the ultimate parent company, NAICS codes, latitude and longitude, and phone number. A number of these fields are proposed as Tier 2 fields in Table 1, as they inform

[30] Due to the variation in Federal agencies and programs, this Workgroup is not currently prescribing a process for agencies to use to implement the Common Data Elements, but a future effort could look into this.

[31] For further information, see "Federal Spending Transparency Data Standards" at MAX.gov, available at: https://max.gov/maxportal/assets/public/offm/DataStandardsFinal.htm

[32] NIEM is "a community-driven, standards-based approach to exchanging information." The U.S. Department of Homeland Security the U.S. Department of Justice, and the U.S. Department of Health and Human Services are the stewards of NIEM. or add a qualifier and data source for "majority"
See: https://www.niem.gov/aboutniem/Pages/niem.aspx https://www.niem.gov/aboutniem/Pages/history.aspx

[33] For example, see:
Krizan, C. J., Statistics on the International Trade Administration's Global Markets Program (September 1, 2015). US Census Bureau Center for Economic Studies Paper No. CES-WP-15-17. Available at SSRN: http://ssrn.com/abstract=2661478 or http://dx.doi.org/10.2139/ssrn.2661478

and enhance matching. A number of agencies do not have information on the ultimate parent or intermediate corporate entities in their data sources, but having this information would allow for hierarchical analysis (e.g. having all establishment identifiers each parent company uses). A third set of data elements are Tier 3 fields as they further support validation: the age of the firm in years, and the number of individuals employed by the firm.

Table 1: Initial Proposal for Common Data Fields

Priority	Field(s)	Field Definition	Examples of Relevant NIEM Core Type/Sub-Properties[34]
Tier 1	Identifiers	As many of the following elements as possible or feasible: • Employer Identification Number (EIN) • D-U-N-S Numbers • Legal Entity Identifier (LEI) • Commercial and Government Entity (CAGE) codes • Other cross-agency identifiers (e.g. UI account number) • Other Non-Confidential, Non-Proprietary Identifiers	**nc:OrganizationType** nc:OrganizationIdentification nc:OrganizationOtherIdentification **nc:IdentificationType** nc:IdentificationID
Tier 1	Legal Name	Legal Name of establishment.	**nc:OrganizationType** nc:OrganizationName nc:OrganizationBranchName
Tier 1	Trade Name	Trade Name, "Operating As" Name, or DBA of establishment.	**nc:OrganizationType** nc:OrganizationDoingBusinessAsName
Tier 1	Establishment Physical Location Address	The address is made up of six standardized components: Street Number, Street Name, and Building/Suite, City, State Code, and ZIP+4 or Postal Code. The address should follow the United States Postal Service's standardized address format (fully spelled out, abbreviated by using the Postal Service standard abbreviations or as shown in the current Postal Service ZIP+4 file). See also: http://pe.usps.gov/text/pub28/28c2_001.htm	**nc:OrganizationType** nc:OrganizationLocation **nc:LocationType** nc:Address **nc:AddressType** nc:AddressFullText
Tier 1	Establishment Mailing Address	The address is made up of six standardized components: Street Number, Street Name, and Building/Suite, City, State Code, and ZIP+4 or Postal Code. The address should follow the United States Postal Service's standardized address format (fully spelled out, abbreviated by using the Postal Service standard abbreviations or as shown in the current Postal Service ZIP+4 file). See also: http://pe.usps.gov/text/pub28/28c2_001.htm	**nc:OrganizationType** nc:OrganizationLocation **nc:LocationType** nc:Address **nc:AddressType** nc:AddressDeliveryPoint
Tier 1	Establishment Physical Location County Code	County codes from US Census and the American National Standards Institute (ANSI INCITS 31:2009). Available at: https://www.census.gov/geo/reference/codes/cou.html	**nc:AddressType** nc:LocationCounty census-3.0.1:USCountyCodeType

[34] From NIEM 3.2 (current release). Available at: https://www.niem.gov/technical/Pages/current-release.aspx
"NIEM core consists of data elements that are commonly understood and defined across domains, such as person, activity, document, location, and item. It's governed jointly by all NIEM domains."
From: https://www.niem.gov/technical/Pages/The-Model.aspx

Priority	Field(s)	Field Definition	Examples of Relevant NIEM Core Type/Sub-Properties[34]
Tier 1	Country Code	Country Code as defined by ISO 3166 (codes for the countries, dependent territories and special areas of geographical interest), FIPS 10-4, or Geopolitical Entities, Name and Codes (GENC).	**nc:AddressType** nc:LocationCountry nc:LocationCountryFIPS10-4Code nc:LocationCountryGENCCode nc:LocationCountryISO3166Alpha2Code
Tier 1	Time Stamp of Collection, Time Stamp when observation Last Edited, and Lag Time.	Date and Time of collection, Date and Time when an observation was last edited, and Amount of Time between the reference period and time the data are available.	**nc:DateRepresentation** **nc:MetadataType** nc:ReportedDate nc:LastUpdatedDate **nc:DateRangeType**
Tier 1	E-mail Address	A corporate E-mail address for respondent.	**nc:OrganizationType** nc:OrganizationPrimaryContactInformation **nc:ContactInformationType** nc:ContactEmailID
Tier 1	Web Site	URL of web site for entity.	**nc:OrganizationType** nc:OrganizationPrimaryContactInformation **nc:ContactInformationType** nc:ContactWebsiteURI
Tier 2	Telephone Number	Telephone number for entity.	**nc:OrganizationType** nc:OrganizationPrimaryContactInformation **nc:ContactInformationType** nc:ContactTelephoneNumber
Tier 2	Other Intermediate Entity Legal Name (If Applicable)	Legal Name of Other Intermediate Corporate Entity.	**nc:OrganizationType** nc:OrganizationName
Tier 2	Other Intermediate Entity Identifier	Identifiers of Other Intermediate Corporate Entity (see above guidance on Identifiers)	**nc:OrganizationType** nc:OrganizationName **nc:OrganizationType** nc:OrganizationIdentification nc:OrganizationOtherIdentification **nc:IdentificationType** nc:IdentificationID
Tier 2	Ultimate Parent Entity Legal Name	Legal Name of Ultimate Parent.	**nc:OrganizationType** nc:OrganizationParent nc:OrganizationParentAffiliate nc:OrganizationParentOrganization **nc:OrganizationType** nc:OrganizationName

Priority	Field(s)	Field Definition	Examples of Relevant NIEM Core Type/Sub-Properties[34]
Tier 2	Ultimate Parent Identifiers	Identifiers of Ultimate Parent (see above guidance on Identifiers)	**nc:OrganizationType** nc:OrganizationParent **nc:OrganizationType** nc:OrganizationIdentification nc:OrganizationOtherIdentification **nc:IdentificationType** nc:IdentificationID
Tier 2	NAICS Code	Six-Digit NAICS Code. 2017 NAICS revision.[35]	--
Tier 2	Latitude and Longitude of Establishment Physical Location Address	Latitude and Longitude of Establishment Physical Location Address, aligned to ISO 6709 and Federal Geographic Data Committee (FGDC.gov) established standards.	**nc:OrganizationType** nc:OrganizationLocation **nc:LocationType** nc:LocationGeospatialCoordinate
Tier 3	Establishment Age	Age of establishment in years.	**nc:DateRangeType** **nc:OrganizationType** nc:OrganizationIncorporationDate
Tier 3	Employment	Count of employees who are on the payroll, ideally for the pay period including March 12 of each year.[36] [37]	--

The fields in Tier 1, Tier 2 and Tier 3 vary in importance for individual offices. The Tiers shown in Table 1 reflect consensus on prioritized fields, but offices should account for variations. For example:

- Tier 2 elements become Tier 1 elements in establishment-based surveys. For example, phone numbers become an important Tier 1 field for not only identifying individual stores within an enterprise with one EIN, but also, having appropriate contact information for each store for future data collection activities. When matching on a city, there are many instances where the postal address is different from the municipal location. For example, New Jersey townships do not relate to what the postal service calls the location. Having additional fields, such as the County Code, further assist in matching.
- Fields in Tiers 2 and 3 may take on particular importance for regulatory activities. For example, it may be critical for a program to have information on intermediate and parent entities in order to engage the proper stakeholders for a compliance activity. Or, if an agency undertakes an investigation of a business that does not have an office, fields related to contact information may become critical for identification purposes.
- Also, some Tier 1 elements, such as the corporate e-mail address (following the "@" section) and/or Web site can be used to some extent as a self-identified definition of the "firm".[38]

[35] http://www.census.gov/eos/www/naics/federal_register_notices/notices/fr08au16.pdf
[36] Derived from: http://www.census.gov/programs-surveys/cbp/about/glossary.html
[37] In instances where an entity has no employees, this field is equal to zero.
[38] For more information, see:
Krizan, C. J., Statistics on the International Trade Administration's Global Markets Program (September 1, 2015). US Census Bureau Center for Economic Studies Paper No. CES-WP-15-17. Available at SSRN: http://ssrn.com/abstract=2661478 or http://dx.doi.org/10.2139/ssrn.2661478

Federal agencies also want standardization and validation approaches for the Common Data Elements shown in Table 1, to maximize the quality of the data and ease of matching across Federal data sources.

- The table references NIEM Core data elements to take advantage of existing data standards, to which a number of Federal agencies have already agreed, for the Common Data Fields.[39]
- Table 1 shows Establishment Physical Location Address, and Establishment Mailing Address, with data entry of the street number, street name, and building/suite in separate variables, to expedite data cleaning. Table 1 also recommends use of the United States Postal Service's standardized address format for these fields.
- The data elements Legal Name, Trade Name, Other Intermediate Entity Legal Name, and Ultimate Parent Entity Legal Name should follow common formatting and validation standards to facilitate matching. For example, agency processes should account for legal entities which have a slightly different legal name in each state. Due to the range of approaches, which depend on project goals, the workgroup is not able to recommend a single standard or method. Rather, agencies should take advantage of existing methods for standardization and validation, which are further discussed in Appendix B. Also, Federal agencies should explore use of resources that help delve into correct names as well as legal filings.[40]

Data Collection Improvements

Federal agencies can achieve potential cost savings in data matching and reuse when they improve the quality of data at the point of entry. They accomplish this with changes in collection methods, subject to any legal requirements (e.g., the Paperwork Reduction Act of 1995), or with validation steps at the time of input. There is a range of methods for improving the quality of fields at the point of collection, resulting in higher quality data, fewer resources expended to clean data, and burden reduction for the regulated community.

First, agencies benefit from ensuring that they have clear reporting guidance for data collection activities to minimize variability in data quality. Robust and clear guidance, combined with a collection tool which reinforces the guidance, ensures that agencies minimize efforts in cleaning data.

Additionally, Application Programming Interfaces (APIs) for validating data have been proven to significantly improve data quality for matching while simultaneously reducing reporting burden. APIs on the front and back ends of data entry are highly useful for validating and standardizing the common data elements shown in Table 1. For example, the U.S. Environmental Protection Agency's Facility Registry Service (FRS), which integrates data on over 4 million establishments and places of interest from across 90 different systems, built an API which took advantage of existing, previously-collected data from internal data sources for entity resolution at the point of data collection.[41] The API allows reporters to identify establishments by searching and retrieving information that was previously reported via other systems from which the FRS ingests data. When a user searches for a particular firm, the application then allows newly reported information to be associated with existing, known establishment information and identifiers. This interface has been used for other EPA programs, such as the Toxic Release Inventory (TRI).[42] Within the TRI program, the EPA has seen significant data quality improvements and has estimated a burden reduction of 140,269 hours for reporters.

---

[39] See also: NIEM 3.2 (current release). Available at: https://www.niem.gov/technical/Pages/current-release.aspx

[40] One example of this is OpenCorporates, which scrapes the legal filings from the various State Departments of State. See: https://opencorporates.com/.

[41] "Facility Registry Service," *US Environmental Protection Agency, Office of Environmental Information*, Retrieved 09 Sep 2016 https://epa.gov/frs

[42] US EPA Toxic Release Inventory https://www.epa.gov/toxics-release-inventory-tri-program

Lastly, agencies benefit from having a strong understanding of respective data sources to be matched, anticipating data quality or consistency issues, and estimating the likely overlap in records for a point of comparison. Many agencies also provide points of contact who can discuss unusual or anomalous results, as well as steps they routinely use to optimize any attempted comparisons using their data. Common practices include:

- accounting for context in databases (such as knowing the context of the letters you are matching in a name),
- ensuring metadata is sufficiently descriptive to differentiate fields, and
- using forms with specialized functionalities (e.g. SmartForm) to validate and standardize data entered.

Authoritative Source

An authoritative source is a data source that provides current, accessible, and authoritative information on all data of a certain type. An authoritative source for data on business existence and characteristics would greatly improve the matching process, and could aid in reducing burden on employers. Such a source would contain validated information confirming existing or closed businesses, and characteristics for businesses such as geographic location, contact information, size, and industry. Additionally, this source would contain data capturing relationships among different levels of corporate and industry structures, and would be able to provide information on at least a quarterly basis on changes to these relationships. Making an authoritative source accessible would drastically improve the ability to validate data and conduct matches. It could also serve as the basis of a universal identifier in the future. Having an authoritative source would be useful for a variety of purposes including, but not limited to:

- **Entity resolution, including identity verification:** Having an authoritative source for cleaning or matching existing data, both during data entry (for example, confirming the identity of a reporting entity) and after (for example, when trying to match two data sets, executing resolution algorithm to show that two entities are actually the same), would reduce the time needed to reconcile data sources when matching them.
- **Sampling:** An authoritative source could enable Federal agencies to build valid sampling frames for surveys or program evaluations without having to rely on proprietary data sources.
- **Administrative purposes:** Authoritative sources could also assist with statistical processes within regulatory agencies, such as creating more accurate or precise estimates of various administrative or economic measures used as inputs for assessing agency performance (for example, examining compliance trends by normalizing for the size of an industry in a local area).

Currently there is no Federal data source that fits this description precisely. Appendix A describes a set of four potential data sets that could be the base for an "authoritative source." There are advantages and drawbacks to each of these sources, and further work would need to be done to determine how to enable Federal agencies to use these as authoritative sources. They include: the Business Register (Census Bureau), the Quarterly Census of Employment and Wages (BLS), the Business Master File (IRS), and the GLEIF Concatenated File (Global Legal Entity Identifier Foundation). While the three government data sources offer the greatest coverage, and are of the highest quality for identified uses, they have many legal and practical requirements and restrictions.[43] [44] The GLEIF Concatenated File has the advantage of being publicly available, but at this moment does not

[43] U.S. Census: Title 13. https://www.census.gov/history/www/reference/privacy_confidentiality/title_13_us_code.html
[44] IRS: Title 26. https://www.irs.gov/tax-professionals/tax-code-regulations-and-official-guidance

cover a range of industries and establishments in a manner similar to the other three sources. Please see Appendix A for additional information.

The Workgroup also suggests further investigation towards creating federal protocols to support data sharing, such as the creation of centralized data sharing guidance. For example, some statistical agencies can allow programs with approved projects (including program evaluations) to detail employees to overcome impediments to cross-agency data sharing. Additionally, some agencies might adopt the model set by the State of Illinois's Department of Innovation and Technology, which in 2016 developed a government-wide MOU that defined a common data taxonomy, standardized internal controls, streamlined data sharing, and established an arbitration process. As a result, data requests are no longer "ad-hoc." They are processed in a matter of days—like FOIA requests—through a process described by the Department's General Counsel as "safe, quick, and transparent."[45] Several agencies, notably the Social Security Administration, have strong internal data sharing protocols that could be replicated on a larger scale. It may also be possible for the Federal government to follow aspects of the Illinois model without requiring changes to statute.

In light of the significant legal barriers, or policy and legal interpretations associated with the confidential sources noted in Appendix A,[46] agencies can instead maximize the use of their existing data by setting up validation and entity resolution checks at the point of data entry, or determine how to improve data sharing practices accounting for the models just described.

Over the long term, Federal Agencies may consider a range of options for additional best practices. For example, it would be beneficial to look into the possibility of combining the authoritative sources noted in Appendix A to create a new mapping table that only contains data necessary for entity resolution, and which could be available to a wider audience of Federal agencies. This could be treated as distinct from the other sources, may not have the same access restrictions as the underlying source, and would be similar to data sets where the data are confidential but can be shared in a masked or aggregated form. Alternatively, it would be useful to further examine to what extent an authoritative source could be constructed that does not contain confidential data, and to examine the quality of such a source relative to confidential sources.

NEXT STEPS

The Workgroup has developed additional recommended topics for investigation that could help improve the sharing of employer data:

- The establishment of an interagency community of practice and repository for sharing methods, code, and approaches to data collection and matching. Federal agencies might benefit from continuing to share knowledge in a structured manner, to ensure that Federal analysts take advantage of the most efficient matching and data collection approaches for a variety of applications. Offices within the Workgroup are able to assist in the establishment and maintenance of the group so that it would continue to be meaningful and useful for participants.
- Consultation with a broader Federal and external audience to gain additional insights from external experts and front-line statisticians.
- The establishment of a centralized data sharing "referee." Federal agencies have expressed an interest in finding ways to expedite and facilitate data sharing, and have identified a centralized office as key to achieving this. Such an office could develop common data sharing protocols, serve as a library of data

[45] Testimony of Michael Basil, General Counsel, Department of Innovation and Technology, State of Illinois before the Second Meeting of the Commission on Evidence-Based Policymaking, September 9, 2016.

sharing agreements, and otherwise serve as a home for knowledge and process for interagency data sharing.

CONCLUSION

There is substantial potential to achieve efficiencies in matching U.S. employer data across Federal data sets for data analysis, evaluations, and statistical activities, based on common needs across agencies. The greatest barrier to matching data on employers across data sets is the lack of a common, or universal, business identifier. Eliminating this obstacle by developing an infrastructure to create, assign, and manage a universal identifier could result in cost savings in matching but would require a major investment of time and taxpayer resources. Assuming that the identifier could be created, it would be a challenge to enforce consistent use of such an identifier by all employers without statutory changes. This identifier would need to capture various corporate/industry levels and change over time, but no Federal entity collects all of this information.

Nevertheless, there are immediate steps agencies can take to adopt best practices for data collection and matching, and there are ways to maximize the use of existing authorities and data sets to enable matching. In particular, leveraging data elements common among Federal agencies which, in combination, can constitute a universal unique identifier, would facilitate efficiencies for matching Federal data sources. The utility of Federal data sources can be increased by including as many cross-agency identifiers as possible. While these fields are useful for matching, it is important to note that identifiers have confidentiality, privacy, and proprietary concerns.

Moving forward, there are potential places to go further to realize long-term improvements, such as developing a roadmap for implementing common data elements in Federal data sources, developing an authoritative source of business existence and characteristics for Federal agencies, developing a Federal community of practice for matching and entity resolution, and establishing a centralized data sharing “referee.”

APPENDICES

Appendix A: Potential Authoritative Sources

The workgroup identified four existing data sets that could serve as the basis for potential authoritative sources of business existence and characteristics across agency missions and program functions (see Table A1 for additional information):

Business Register, U.S. Census Bureau
The Census Bureau's Business Register contains establishments of all domestic businesses (except agriculture, forestry, fishing, hunting, rail transportation, the U.S. Postal Service, elementary and secondary schools, colleges and universities, labor organizations, political organizations, religious organizations, public administration, and private households) and organizational units of multi-establishment businesses. A single-unit enterprise's primary identifier is its Employer Identification Number (EIN). A unique employer unit identification number identifies each establishment owned by a multi-unit enterprise on the Business Register.

Advantages: The Business Register (BR) covers more than 160,000 multi-establishment companies, representing 1.8 million affiliated establishments, 5 million single establishment companies, and nearly 21 million non-employer businesses (note, Census maintains a separate register for employers and non-employers).[47] The Business Register is updated continuously, and the update frequency varies by its sources. Lags also vary, by the reference period of the sources. The Business Register is one of the most complete, current, and consistent source of establishment- based information about U.S. businesses, and is essential to assuring full coverage and high quality in Federal economic statistics programs.[48]

Considerations: Users should note that the source excludes a significant number of industries, including, most notably, agriculture, education, and public sector. Additionally, due to the lags in the data, users should exercise care in analyses of industries and sectors with high turnover or conversion rates. This data set contains links between establishments and their parent firms, but these links are sometimes recorded a year or more before or after the reference date. [49] Also, the Census Bureau's Business Register is constructed using comingled confidential tax information and non-tax data from various sources, including Census surveys. The fact that the Census Bureau's Business Register contains confidential tax information prevents the Census Bureau from completely sharing its Business Register with other agencies not authorized to receive the confidential tax information under Title 26.

Business Master File, Internal Revenue Service
The Business Master File (BMF) contains data for all Federal business tax returns that meet IRS filing requirements. The data set consists of individually filed returns for a single establishment and consolidated filed returns consisting of a group of related (affiliated) establishments. BMF data are updated continuously but usually become available weekly. The unit of analysis is tax return-based, as filed by the taxpayer; enterprises are not aggregated by the IRS.

Advantages: The data set consists of total population data based on taxpayer filings. Also, the availability of Employer Identification Numbers allows for direct linkages.

[47] See also: https://www.census.gov/econ/overview/mu0600.html
[48] Ibid.
[49] See also: www.bls.gov/osmr/pdf/st140030.pdf

Considerations: The unit of observation is a tax return, which is a higher, and different, level than establishment or physical location for multi-establishment businesses. This differs from both the Census Bureau's Business Register and Quarterly Census of Employment and Wages Business Register. Even for businesses that are not filing within a consolidated group, a return may not represent an establishment, for example in cases where a single business has operations located at more one physical address. Entity information such as addresses is based on taxpayer-reported information and may not necessarily be the actual physical address for matching purposes.

U.S. Bureau of Labor Statistics-Office of Employment and Unemployment Statistics, Quarterly Census of Employment and Wages (QCEW) Business Register (BR)

The QCEW-BR contains employment, wages, and administrative data (name, location, etc.) for over 9.5 million establishments, covering approximately 98% of all employment. The Quarterly Census of Employment and Wages (QCEW) provides data for establishments on monthly employment, total quarterly wages, the number of business establishments, and other business identification information such as address, industry, and federal employer identification number, etc. In addition to being very comprehensive and accurate data, the QCEW-BR is timely: data are available 6 months after the reference quarter making this the most current source of comprehensive business establishment data available.

The QCEW data are the product of a federal-state cooperative program. The data are derived from summaries of employment and total pay of workers covered by state and federal unemployment insurance (UI) legislation and provided by State Workforce Agencies (SWAs). States prepare a microdata file each quarter and submit that to BLS within 15 weeks of the end of the quarter. QCEW data are developed for the 50 states, the District of Columbia, Puerto Rico, and the U.S. Virgin Islands. The summaries are a result of the administration of state unemployment insurance programs that require most employers to pay quarterly taxes based on the employment and wages of workers covered by UI. Employment and wage data for workers covered by state UI laws are compiled from quarterly contribution reports submitted to the SWAs by employers. For federal civilian workers covered by the Unemployment Compensation for Federal Employees (UCFE) program, employment and wage data are compiled from quarterly reports submitted by four major federal payroll processing centers on behalf of all federal agencies.

BLS sets quality standards in state cooperative agreements and provides conceptual, technical, and procedural guidance to the states and uses standardized procedures to process these data and ensure consistent quality across states. State workforce agencies are responsible for collecting the administrative records from their state unemployment insurance system and transform these records into meaningful economic data. In addition to state data quality improvements, BLS conducts additional data review at its regional and national offices to ensure quality.

BLS also enhances the data with two supplemental surveys:

- The Annual Refiling Survey, which allows BLS and the states to collect updated North American Industry Classification System (NAICS) industry codes, geographic county codes, and address information for business establishments.
- The Multiple Worksite Report, which allows BLS to collect detailed monthly employment and total wages each quarter for businesses with more than one location. This allows the program to capture business births and deaths in a timely and frequent manner and accurately capture changes in ownership as a result of mergers and acquisitions.

BLS links the microdata for each business establishment across quarters to create a longitudinal record. These data are available starting in 1990 through the most recent data available: second quarter 2016. This linked

microdata file serves as the sampling frame for BLS establishment-based surveys such as the Current Employment Statistics (CES), a key survey used for the publication of the monthly Employment Situation. Other BLS programs that use the QCEW microdata for sampling purposes include the Job Openings and Labor Turnover Survey (JOLTS), Occupational Employment Statistics (OES), Producer Price Index (PPI), Survey of Occupational Injuries and Illnesses (SOII), National Compensation Survey (NCS), including the Employment Cost Index (ECI) and Employer Benefits Survey (EBS), and the new Occupational Requirements Survey (ORS). The Local Area Unemployment Statistics (LAUS) program also uses the QCEW as its source of employment when CES estimates are not available.

In addition, BLS publishes Business Employment Dynamics (BED) statistics drawn from this linked microdata set. BED statistics are created from the linked individual business establishment records that are tabulated to create aggregate time series for national and local business establishment openings, closings, expansions, and contractions, all by industry. Over the years, additional variables have also been created, including establishment age, survival rates, and firm size.

The QCEW program publishes tabular data for the nation, states, metropolitan statistical areas, and counties at a detailed NAICS 6-digit industry level about 6 months after the end of the reference quarter. The release date for QCEW data has been moved up by a total of 3 weeks since 2012. The data are valued for their comprehensiveness, accuracy, relevance, and timeliness.

Business Employment Dynamics statistics are published in the month following the release of the QCEW tabular data. These statistics have gained a wide user base as economists and other analysts continue to examine the role of employment dynamics in the U.S. economy. Key user groups of BED statistics include the Federal Reserve System, the Small Business Administration, and academics.

Key users of tabular QCEW data include Congress, state and local economic development agencies, state revenue forecasters, and numerous Federal agencies. Over $300 billion in public funds are allocated based on QCEW data.

Among the many users of QCEW data, four demonstrate some of the data's varied roles:

- The Employment and Training Administration (ETA) uses QCEW data to measure the solvency of unemployment insurance trust funds and to develop the statistical adjustment models for measuring quarterly performance of Workforce Innovation and Opportunity Act (WIOA) funded core programs.
- The Bureau of Economic Analysis (BEA) uses the quarterly QCEW data to develop county, state, regional, and national personal income estimates, a component of the gross domestic product, and to conduct related statistical research and analysis. In 2015, covered workers received $7.385 trillion in pay, representing 94.0 percent of the wage and salary component of personal income and 40.9 percent of the gross domestic product.
- Census Bureau
    - Since 1990, BLS has shared NAICS codes with the Census Bureau to improve industry coding consistency and reduce respondent burden and costs. The Census Bureau uses these industry data in its Business Register, which serves as a source of sampling frames for frequent business surveys (such as the Annual Survey of Manufacturers) and as a basis for statistical tabulations. The most important benefits of this data-sharing project are relieving American businesses of unnecessary response burden, improving industry coding for the Census Bureau, improving usability and promoting consistency between federal statistical products, and reducing redundancy between agency statistical programs, to the exceptional benefit of the American taxpayer.

    - In addition, after states and BLS have edited and curated the microdata, the QCEW data file is sent to the Census Bureau, where the data serve as a primary input for the Longitudinal Employer-Household Dynamics (LEHD) program.
    - QCEW data are also used to calibrate the joint BLS-Census Current Population Survey after each decennial census.
- Outside researchers also apply for access to the QCEW microdata in a protected environment for projects that are relevant to the mission and scope of BLS.

By publishing QCEW and BED data, and by sharing these data amongst our statistical partners, BLS provides informational infrastructure that enhances the ability of the public and private sectors to make evidence-based decisions.

Advantages: The QCEW-BR contains data for all industries (including government), and coverage is mandatory, therefore compliance is high. In addition, the data set has consistent terms, is accurate, timely, relevant, and has a strong validation process. The QCEW-BR is sharable, subject to state data sharing restrictions. It is also a high frequency data set with quarterly data collection for monthly employment values. The QCEW data set is sustainable and scalable. There are a number of matching projects and new data products that can be developed with little or no new response burden at little or no additional cost (i.e., nonprofit data, etc.).

Considerations:
Exclusions from QCEW include self-employed workers, most agricultural workers on small farms, all members of the Armed Forces, elected officials in most states, most employees of railroads, some domestic workers, most student workers at schools, and employees of certain small nonprofit organizations.

There are some states who do not agree to share their data. A law change allowing for full data sharing would enhance the usability of QCEW data.

Global Legal Entity Identifier Foundation, Concatenated Data File
The Global Legal Entity Identifier Foundation (GLEIF) publishes the updated GLEIF Concatenated File daily. This file contains the content of the individual files, published by the Legal Entity Identifier (LEI) issuing organizations, which list all LEIs issued to legal entities and related LEI reference data. The data provides information on a legal entity identifiable with an LEI.

Advantages: The key advantage of the GLEIF Concatenated File, relative to the other sources noted, is that it is publicly available. The Global Legal Entity Identifier Foundation notes on their web site, "The drivers of the LEI initiative, i.e. the Group of 20, the [Financial Stability Board] and many regulators around the world, have emphasized the need to make the LEI a broad public good. The Global LEI Index, made available by GLEIF, greatly contributes to meeting this objective. It puts the complete LEI data at the disposal of any interested party, conveniently and free of charge."[50] The GLEIF Foundation will also be collecting information on parent and subsidiary entities during the annual re-registration for LEI numbers in 2017, and plans to continue this data collection in future years.[51]

[50] "Introducing the Legal Entity Identifier (LEI)." Global Legal Entity Identifier Foundation. <https://www.gleif.org/en/about-lei/introducing-the-legal-entity-identifier-lei>. Accessed 12/9/2016.
[51] https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-2-data-who-owns-whom

Considerations: While the LEI is intended to be a universal and open identifier, coverage is currently low. As of the time writing this there are fewer than 500,000 LEIs issued globally, though this number is expected to increase over time as government's and institutions mature their processes using this identifier. Additional considerations regarding application of the LEI can be found in the November 2015 *Progress report by the Legal Entity Identifier Regulatory Oversight Committee (LEI ROC).*[52]

[52] Available at: https://www.leiroc.org/publications/gls/lou_20151105-1.pdf

Table A1: Options for Authoritative Sources

Component Agency or Office	Data set Name	Purpose of the Data Collection	Access Restrictions, Update Frequency and Lags	Coverage	Definition of "Company"	Unit of Analysis, Corporate Structure and Relationships	Quality of Fields for Matching
U.S. Census Bureau	**Business Register**	To provide a current and comprehensive database of U.S. business establishments and companies for statistical program use.	The Business Register information is confidential [Title 13 and Title 26, US Code]. Access is restricted to persons specially sworn to uphold the confidentiality provisions of Title 13 and Title 26. Data are updated continuously, update frequency varies by sources; lags vary by the reference period of the sources.	Establishments of all domestic businesses (except agriculture, forestry, fishing, hunting, rail transportation, the U.S. Postal Service, elementary and secondary schools, colleges and universities, labor organizations, political organizations, religious organizations, public administration, and private households) and organizational units of multi-establishment businesses. The Business Register (BR) covers more than 160,000 multi-establishment companies, representing 1.8 million affiliated establishments, 5 million single establishment companies, and nearly 21 million non-employer businesses.	An establishment is a single physical location where business transactions take place and for which payroll and employment records are kept. Groups of one or more establishments under common ownership or control are enterprises. A single-unit enterprise owns or operates only one establishment. A multi-unit enterprise owns or operates two or more establishments.	Establishment-based.	The Business Register is one of the most complete, current, and consistent source of establishment- based information about U.S. businesses, and is essential to assuring full coverage and high quality in Federal economic statistics programs. Examples of quality considerations for this source include: - The annual Company Organization Survey covers 30 percent of multi-unit companies and a small sample of firms that were single-unit firms in the most recent quinquennial Economic Census, so establishment openings and closings in the firms not covered may not be reflected in the business register until after the next Economic Census (though the Census Bureau takes measures to address this). - The business register is divided into employer and non-employer business registers based on payroll employment. Some firms lease their employees from Professional Employer Organizations (PEOs) or use independent contractors. Such firms may appear in the non-employer business register despite having large revenues and many leased and/or contract employees. (https://www.census.gov/econ/overview/mu0600.html)

U.S. Bureau of Labor Statistics-Office of Employment and Unemployment Statistics (OEUS)	**Quarterly Census of Employment and Wages (QCEW) Business Register (BR)**	To provide a quarterly census of all establishments under State unemployment insurance programs, representing about 98 percent of employment on nonfarm payrolls. This database of U.S. business establishments serves as the basis for multiple statistical programs; Sampling frame & benchmark(CES, JOLTS, PPI, OES, LAUS, SOII, and NCS, which includes the ECI, EBS and ORS), labor market research, Business Employment Dynamics (BED) data.	All microdata are confidential subject to BLS non-disclosure standards. The QCEW BR is updated quarterly, and the data become available 6 months after the reference cycle.	Employment, wages, and administrative data (name, location, etc.) for over 9.5 million establishments covering approximately 98% of all employment.	An economic unit that produces goods or services, usually at a single physical location, and engages in one or predominantly one activity. --Potential lay synonyms: business, worksite, brick& mortar, site, storefront. Lay users my use "establishment" interchangeably with "firm" In the QCEW BR, however, there is a significant distinction between the two terms.	Establishment-based. Multi-unit enterprises may use multiple EINs, and they may use different ones when reporting to different Federal agencies, complicating the matching process. The data set indicates whether an establishment is a single or multi-location establishment.	In order to ensure the highest possible quality of data from the QCEW program, BLS and the States verify and update, if necessary, the NAICS, location, and ownership classifications of all units on a 3–year cycle. Government units in public administration are not reviewed routinely.

Internal Revenue Service - Statistics of Income	**Business Master File**	IRS: The purpose of data collection is mainly for determining Federal tax liability for businesses required to file. SOI: The data are used to produce statistics on income, deductions, credits and other taxes, as reported by businesses. The current design is a probability sample stratified by Business type (as indicated by the IRS form filed) and either by size of total assets alone or size of total assets and a measure of income.	Federal Tax Information (FTI) is confidential [Title 26, US Code] and shared with other government agencies under IRC 6103(j) provisions. Data are updated continuously; Data becomes available weekly.	IRS: All Federal business tax returns that meet IRS filing requirements. SOI: Selected active Federal business tax returns based on SOI's sample design.	Definition of "company" is based on Title 26 requirement.	Tax Return. The data set consists of individually filed returns which can represent: corporations (a single establishment; subsidiary establishment or consolidated filed returns representing a group of establishments); partnerships and other pass-through entities; or sole proprietorships.	Most matching is accomplished using the EINs provided by entities as reported on Federal tax returns. SOI uses exact matching with EINs and data processing begins with information already extracted for IRS administrative purposes. SOI performs limited internal "data cleaning" for statistical purposes. This includes organizing data to make it structurally consistent, coding data items to make them analytically useful, and validating values to ensure mathematical consistency. Contact information is validated as part of routine administrative processing of returns at the time they are received by the IRS.
Global Legal Entity Identifier Foundation	**GLEIF Concatenated File**	The Global Legal Entity Identifier Foundation (GLEIF) publishes the updated GLEIF Concatenated File daily. This file contains the content of the individual files, published by the Legal Entity Identifier (LEI) issuing organizations, which list all LEIs issued to legal	Publicly available without restrictions	All legal entities are eligible to receive an LEI. As of the time writing this there are 465,397 LEIs issued globally.	As defined in ISO 17442, the standard underlying the Legal Entity Identifier (LEI), the term 'legal entity' includes, but is not limited to, unique parties that are legally or financially responsible for the performance of financial transactions or have the legal right in their jurisdiction to enter independently into legal contracts, regardless of whether they are incorporated	Legal entity based	Very high quality where available. There is a validation effort done whenever an entity applies for an LEI. This ensures a distinct unique record exists for each identifier (). The LEI identifies an entity across multiple data sets where required by regulation. See the following report for additional information: https://www.leiroc.org/publications/gls/lou_20151105-1.pdf

		entities and related LEI reference data. The data provides information on a legal entity identifiable with an LEI			or constituted in some other way (e.g. trust, partnership, contractual). It excludes natural persons, but includes governmental organizations and supranationals.		

APPENDIX B: Best Practices for Matching Data

The Workgroup's Methods Inventory yielded a variety of strategies in data collection and data integration which Federal agencies use to work around issues in matching and entity resolution. See Appendix E for specific code examples.

DATA COLLECTION

Federal agencies maximize the use of identification approaches during the data collection phase to improve matching to prevent downstream challenges in matching and entity resolution. These approaches focus on identifying the right entity and industry or corporate relationships, and validating geographic and industry information.

Identifying the right entity, right level and right relationship

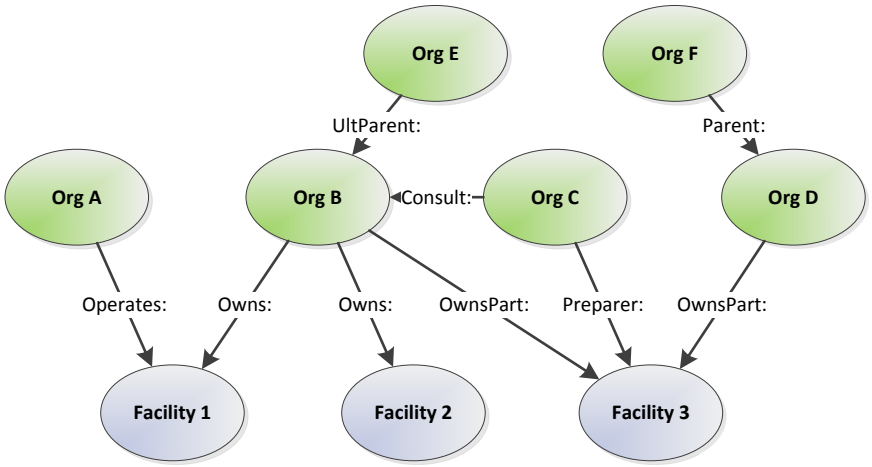
Relationships

Federal agencies often encounter complex industry or corporate structures, and work to identify the correct entities, corporate and industry levels, and corporate and industry relationships at the point of data collection to avoid challenges in matching or entity resolution later.

- For example, a facility may be owned by one company but another may control its operation (this was the case with the Deepwater Horizon disaster – the rig was owned by Transocean, but leased by British Petroleum (BP) and operated by Transocean and other contractors under BP's direction[53]).
- There may also be cases where a facility has one or more contractors who each have their own compliance, regulatory and reporting requirements co-located within the same facility, such as a steel mill.
- Another scenario exists in the case of joint ventures, where a facility may be owned and/or operated by multiple parties. Additionally, there may be an owner or operator entity being captured at the facility level, but also some degree of org parent information, whether an ultimate global parent or domestic parent.
- Lastly, Figure B1 shows an agency example where Facility 1 is owned by Organization B, but operated by Organization A; Organization B also owns Facility 2 and is part owner (via Joint Venture or other vehicle) of Facility 3, Organization B lists as its ultimate parent Organization E, however there may be intermediate entities not captured; Organization B also retains Organization C as a consultant to aid in preparation of regulatory documentation for Facility 3. Facility 3's remainder owner is Organization D which lists Organization F as its parent, however Organization F might not be its ultimate parent.

[53] See also: Final Report of the National Commission. http://oscaction.org/resource-center/commission-reports-papers/

Figure B1: Example of a Complex Industry Structure at the point of Data Collection



It is due to cases such as these that applications and data owners need to be cognizant of the importance of very robust and clear guidance, and good data structures for capturing this information, paying close attention to the contextual relationships. Agencies have found that relational, NoSQL, RDF triple stores and/or graph databases work well for capturing these types of complex relationships in the data.

Establishment Facility Name

Often, difficulties in matching stem from inconsistent approaches and guidance for providing establishment names, for example providing only one organization name for a campus which comprises multiple operations and facilities which report to agencies separately from each other (for example a large university which may have physical plant and utilities versus labs and other facilities each with independent permitting, compliance and reporting responsibilities). In some cases, one may need sufficient information in order to disambiguate, requiring enough information and consistency in reporting. In other instances, one may need to merge and de-duplicate data, and a lack of consistency can also be an issue in achieving this.

Table B1: Variations in Establishment Name

Establishment Name
Widgetco
WidgetCo Plant 2
Plant 2
Second Unit Widget Co

Contact Information

Data elements such as email addresses or telephone numbers can also be used to identify relationships in data and assist in entity resolution; for example one single point of contact is associated with over 4,000 chain drugstores in EPA's Resource Conservation and Recovery Act data set. Similarly, telephone numbers can also be a useful tool in entity resolution. As with the prior cases, data entry and reporting interfaces should include

validators to ensure proper formatting and values that appear valid. A number of JavaScript form validators exist [54] which can aid in these tasks.

Web sites, and e-mail addresses (after the "@" section) can not only serve as potentially valuable linking variables, but also as firm identifiers. Web sites and e-mail addresses are not subject to the same sorts of variation or mis-spellings that corporate names are. Similarly, while a corporation may have many regional telephone numbers, it will only have one domain name. Finally, companies will self-sort their economic activities into the appropriate Web domain names, likely based on activity, in cases where they undertake multiple, disparate business activities.

Geography-Related Fields

Address fields

As with entity reporting, for address fields, clear guidance and documentation needs to be implemented for collection interfaces, as well as providing the appropriate data structures for capturing separate values for physical and mailing/administrative address.

Ideally, systems should provide geocoding capabilities which can validate and standardize entered street addresses, i.e. 528 South Fourth Street standardized to 528 S 4th St per USPS standard to aid in disambiguation. For large factories, in particular, it is also useful to specify where on the establishment the latitude and longitude will be established. Geocoding APIs typically also provide an effective mechanism for parsing and standardizing street address elements, such as house number (528) / street directional (S) / street name (4th St). Additionally, a geocoding API[55] can provide latitude/longitude values which can easily be used to display a web map view for additional visual verification of site locations in reporting interfaces.

There are also cases where establishments might not have conventional street addresses, for example remote facilities in oil and gas sectors, or many Puerto Rico addresses which are linear addresses by distance marker along a route. Some typical cases may involve public works infrastructure where often the City Hall address is provided, or a PO Box is provided in place of the physical address, when quite likely both are wanted. Many systems also struggle to differentiate adequately in the case of multi-establishment addresses, whether office buildings, suites, incubators or industrial parks sharing an address, or other similar cases. Foreign addresses are also often problematic in many systems. Table B2 contains examples of commonly-encountered but difficult-to-resolve addresses.

Table B2: Examples of Common, Difficult-to-Resolve Addresses:

Street Address 1	City	State	ZIP
PR 181, KM. 18.2, BO. ESPINO	SAN LORENZO	PR	00754
NW1/4 S27 T48N R77W	JOHNSON COUNTY	WY	99999
H.C. 64 BOX 204	MCFADDEN	WY	82083

[54] "Validators," *formvalidation.io*, Retrieved 09 Sep 2016
http://formvalidation.io/validators/

[55] "Geocoding", *Wikipedia*, Retrieved 09 Sep 2016
https://en.wikipedia.org/wiki/Geocoding

47 MI W OF LARAMIE ON I-80 EXIT 267	LARAMIE	WY	82070
43 32' 11" N 108 49' 29"W	KINNEAR	WY	82516
90TH SUPPORT GROUP, DEV BLVD 320	CHEYENNE	WY	82005

When conventional street addresses cannot be provided, where appropriate and feasible, a clear mechanism should be provided for capturing alternative addressing schemes to improve how these can be dealt with and resolved. For example, WGS84 lat/long values are relatively easy to capture and collect with the advent of ubiquitous GPS technology on mobile devices, as well as the ease of embedding web based mapping applications which can capture a coordinate. When capturing data via latitude/longitude, existing data standards should be implemented, such as ISO 6709:2008[56] for entry and representation of latitude/longitude values. Additionally, consideration should be given to decimal degrees and precision for entry of latitude/longitude. Public Land Survey System (PLSS) references such as NW1/4 S27 T48N R77W are tied to Bureau of Land Management Survey Grid and BLM has web services for determining location based on the descriptive elements[57]. Guidance and consistency in entry of the elements to facilitate parsing for a web service would be another necessary consideration in implementing PLSS entry.

As a final caveat, it should also be noted that postal municipality might not be the same as the jurisdictional municipality where the establishment is located.

Latitude / Longitude Fields

Latitude/Longitude, where available, is a useful data element to use for proximity matching. Degree of precision of the entered latitude/longitude data impacts spatial resolution - and while the distance spanned by a unit of measure will typically be consistent for north/south parallels of latitude, it will vary for east/west meridian values as a function of varying latitude due to meridian convergence at the poles, as shown in Figure B2:

[56] "Catalog: ISO 6709:2008", *International Standards Organization*, Retrieved 09 Sep 2016 http://www.iso.org/iso/catalogue_detail.htm?csnumber=39242

[57] "GeoCommunicator Services," *Bureau of Land Management*, Retrieved 09 Sep 2016 http://www.geocommunicator.gov/geocomm/services.htm

Figure B2: Meridian Convergence

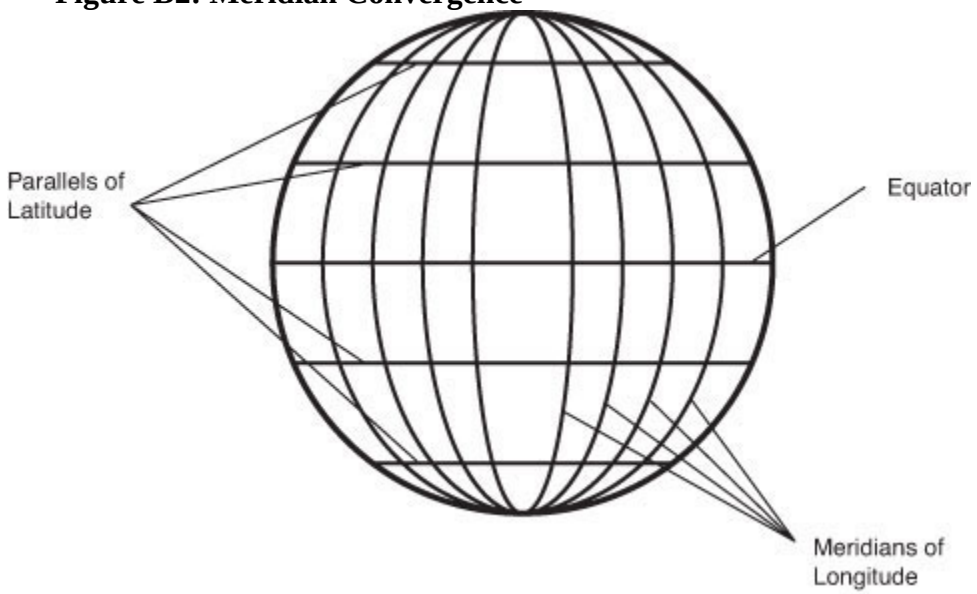


Image source: National Ocean and Atmospheric Administration)

Typical distance values corresponding with various levels of angular precision are as follows:

Table B3: Typical Distance Values Corresponding with Various Levels of Angular Precision

Decimal places	Decimal degrees	DMS	Qualitative scale that can be identified	N/S or E/W at equator	E/W at 23N/S	E/W at 45N/S	E/W at 67N/S
0	1.0	1° 00′ 0″	country or large region	111.32 km	102.47 km	78.71 km	43.496 km
1	0.1	0° 06′ 0″	large city or district	11.132 km	10.247 km	7.871 km	4.3496 km
2	0.01	0° 00′ 36″	town or village	1.1132 km	1.0247 km	787.1 m	434.96 m
3	0.001	0° 00′ 3.6″	neighborhood, street	111.32 m	102.47 m	78.71 m	43.496 m
4	0.0001	0° 00′ 0.36″	individual street, land parcel	11.132 m	10.247 m	7.871 m	4.3496 m
5	0.00001	0° 00′ 0.036″	individual trees	1.1132 m	1.0247 m	787.1 mm	434.96 mm
6	0.000001	0° 00′ 0.0036″	individual humans	111.32 mm	102.47 mm	78.71 mm	43.496 mm
7	0.0000001	0° 00′ 0.00036″	practical limit of field surveying	11.132 mm	10.247 mm	7.871 mm	4.3496 mm

Latitude/Longitude comparison will be discussed later in this document under Haversine Distance matching.

Industry Fields

NAICS/SIC Codes

North American Industry Classification (NAICS)[58] and Standard Industrial Classification (SIC)[59] Codes, may also be a helpful asset toward matching or disambiguating establishment records. For example, certain NAICS and SIC codes are relatively rare, such as Nuclear Electric Power Generation - NAICS: 221113. Additionally, groupings of related and unrelated NAICS/SIC codes could be used in business rules, such as codes which typically do not appear within the same establishment. SIC codes have technically been retired, however some agencies and data sets still use SIC, i.e. SEC, OSHA and others. One thing to note is that NAICS codes are revised on a regular cycle of 5 years, the most recent revisions being 2012 and 2007. As such, in capturing NAICS data, it would be useful to also reference the NAICS version being used, i.e. NAICS:2012. The U.S. Census Bureau provides reference data on NAICS codes, which can be used to populate API-based entry or picklists as opposed to allowing manual entry of codes, to reduce data entry errors.

Hierarchy Field

Knowing which establishments are part of which parent companies (updated quarterly) is incredibly useful for matching efforts at different employer levels.

[58] "North American Industry Classification System, " *US Census Bureau*, Retrieved 09 Sep 2016 http://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2012

[59] "SIC Division Structure," *Occupational Safety and Health Administration*, Retrieved 09 Sep 2016 https://www.osha.gov/pls/imis/sic_manual.html

COLLECTION METHODS

Leveraging Existing Data via API for Entity Resolution: Example case at US EPA

Existing, previously-collected data may be one of the best tools available for entity resolution at the point of data entry. It is critical to highlight examples that enable agency employees to make an internal case for changes. Agencies should note that EPA has begun leveraging its Facility Registry Service (FRS)[60] for that purpose. FRS integrates data on over 4 million establishments and places of interest from across 90 different systems via master data management and a combination of algorithmic methods and manual data steward curation. The FRS team developed an API to improve integration at the point of data collection, by allowing reporters to identify establishments by searching and retrieving information that was previously reported via the various systems that FRS ingests data from.

This capability is illustrated in the following screen shots, in which a user searches for "Finch Paper" on Glen Street in Glen Falls, NY:

[60] "Facility Registry Service," *US Environmental Protection Agency, Office of Environmental Information*, Retrieved 09 Sep 2016 https://epa.gov/frs

Figure B3: ***Entity Search to call API***

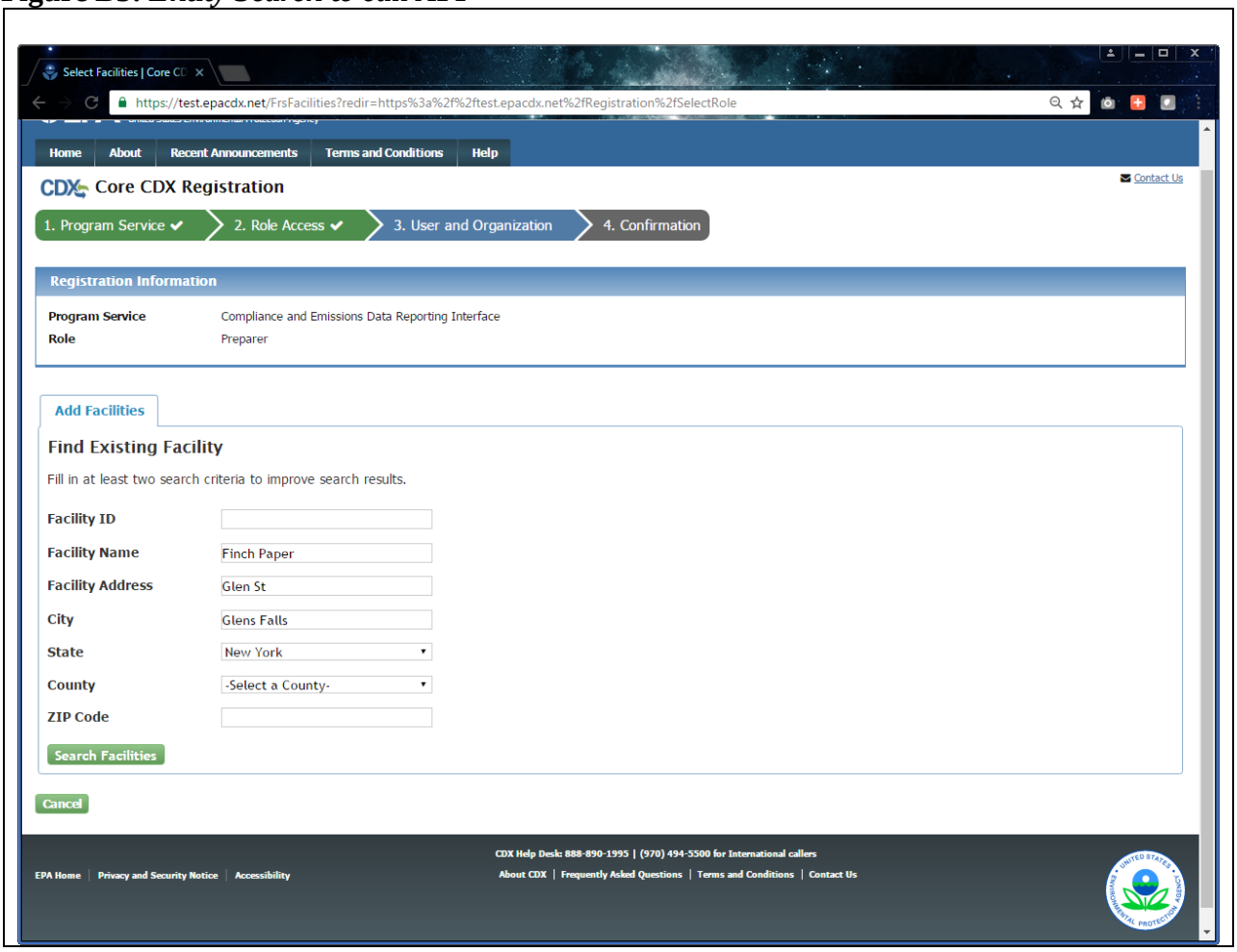
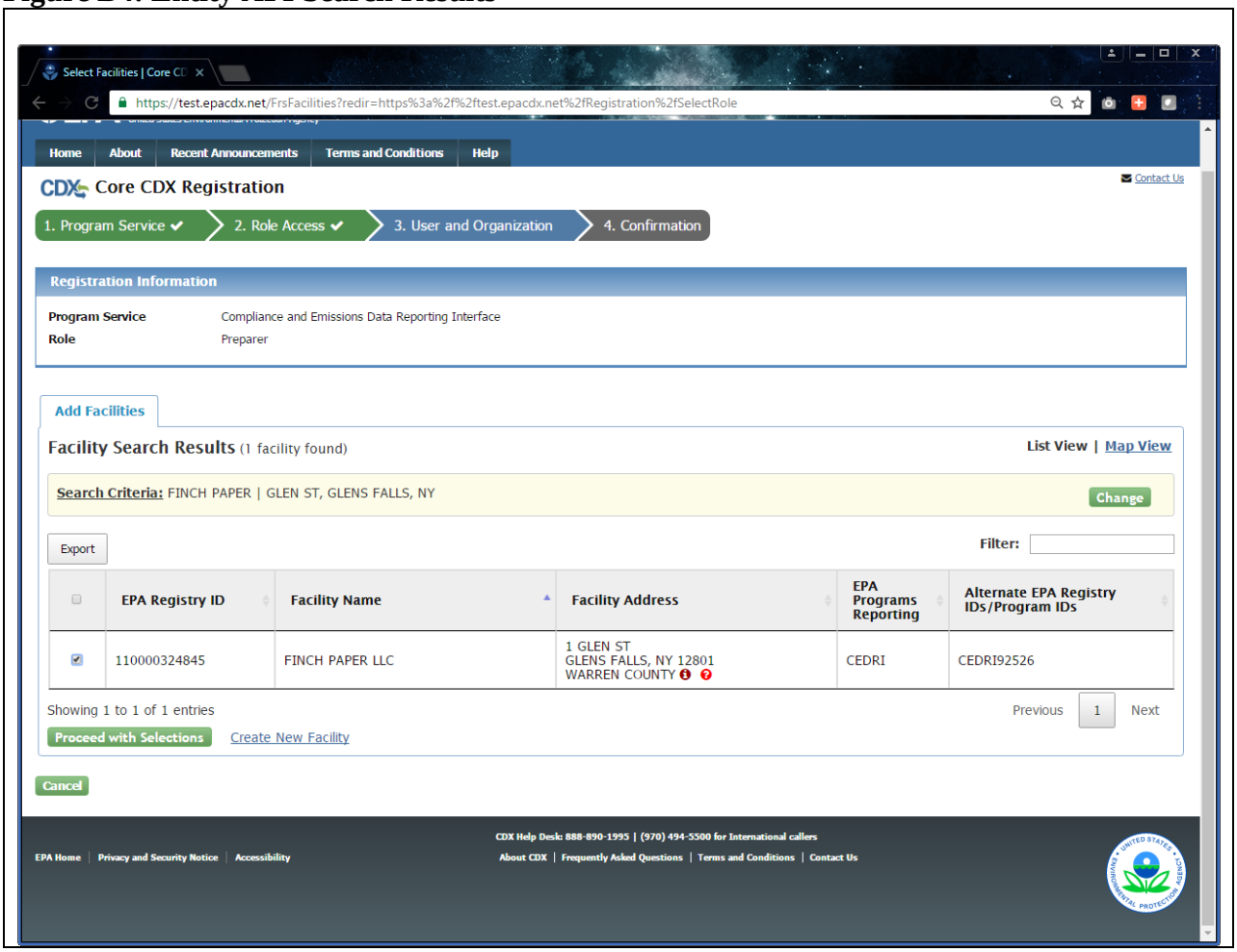


Figure B4: Entity API Search Results



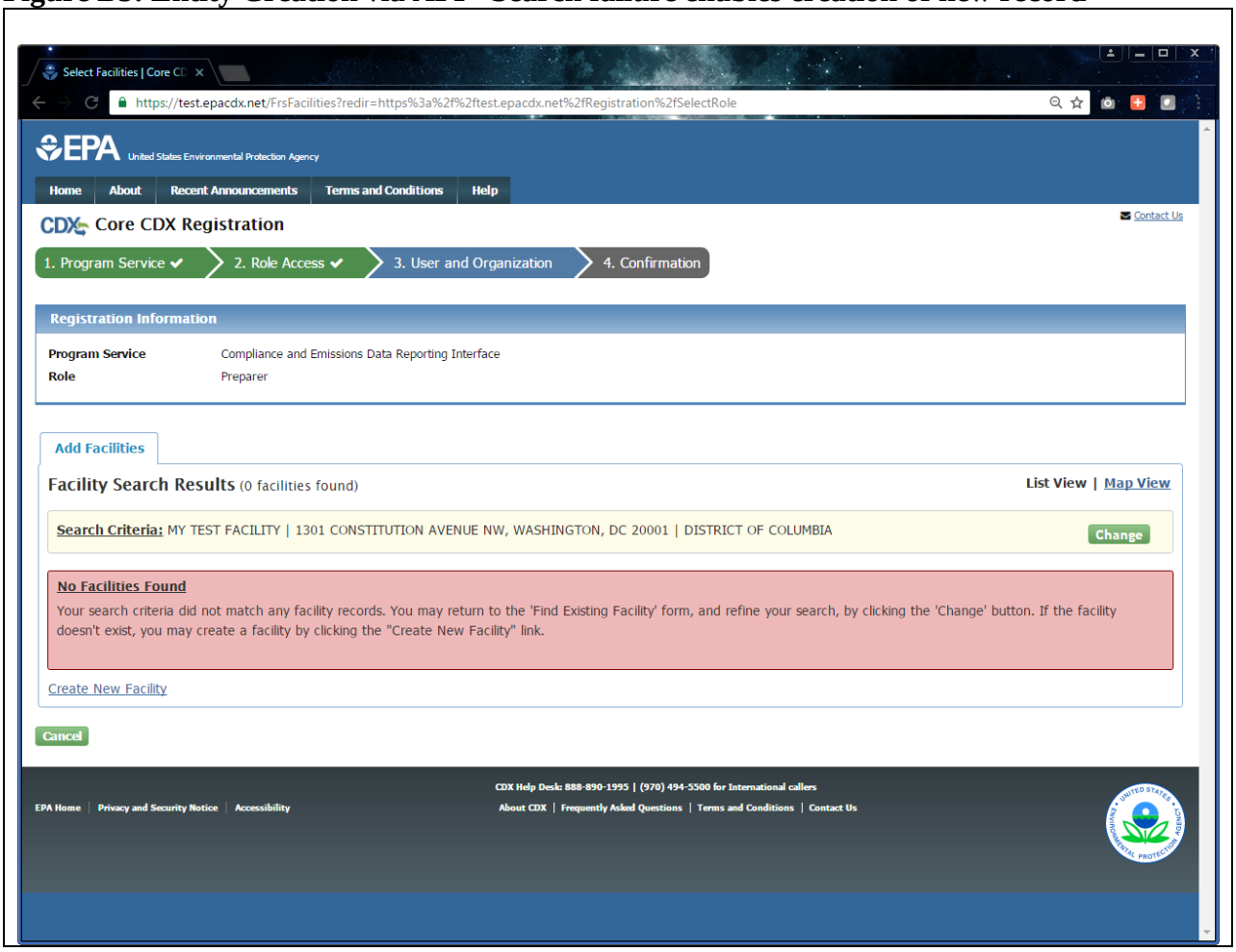
The application would then allow newly reported information to be associated with existing, known establishment information and identifiers.

Initially this API was a read-only API for data retrieval, however it has since evolved to provide additional capabilities, including:

- The ability to submit a suggested update (for example, if an establishment had a typo in its information or if it was acquired by a new firm and the establishment's name needs to change), or
- if a new facility is being reported, the API will allow a master record and unique identifier to be generated in real-time, which will then subsequently be available for other queries.

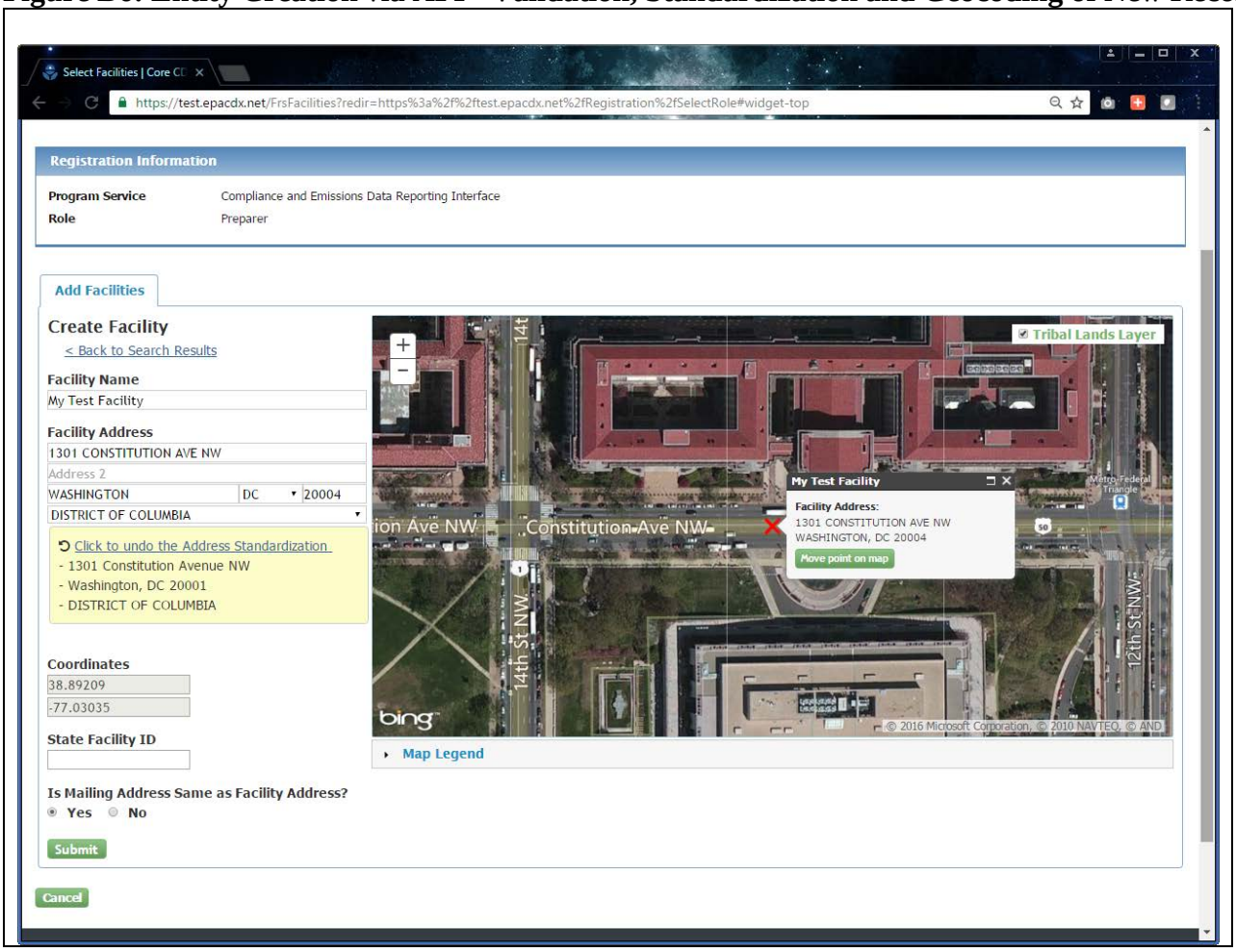
An example of creating a new facility is shown in the following figure:

Figure B5: Entity Creation via API - Search failure enables creation of new record



On clicking the “Create New Facility” button, the user is then presented with a view that presents the information that has been entered, which is standardized and geocoded.

Figure B6: Entity Creation via API - Validation, Standardization and Geocoding of New Record



This reporting interface component is built on an API and has been componentized for reuse across EPA programs, and in the last two years has been integrated into multiple reporting systems, such as the Toxic Release Inventory (TRI), TSCA Chemical Data Reporting (CDR), Compliance Emissions Data Reporting Interface (CEDRI) and others. Within the TRI program, they have seen significant data quality improvements and have estimated substantial burden reductions (as further described in the white paper).

Data Preparation

Prior to attempting matching, some basic exploratory data analysis is recommended to identify the types of data quality challenges noted in the main body of the white paper.

Data sets to be matched may be scattered across different systems, different architectures, and disparate tables, and that preprocessing may need to take place in order to extract the data for analysis, via Extract-Transform-Load (ETL) processes or in the case of loading to systems designed for big data analytics such as HDFS/MapReduce, Extract-Load-Transform as appropriate[61].

[61] Davenport, Robert J. (2008), "ETL vs. ELT: A Subjective View", Retrieved Sep 12 2016 http://www.dataacademy.com/files/ETL-vs-ELT-White-Paper.pdf

It should also be considered that within a data set, there may be substantial turnover of establishments (20% or more within some sectors over a 5-year timespan), via acquisitions, mergers, business startups and closures which can affect name matching; as such, timestamps and windowing within timeframes may be necessary.

Cleansing to Address Inconsistent Data Formats

In cleansing records for entity matching, some key processes include converting all characters to the same case, and removing special characters, extraneous punctuation and extraneous white space between words.

Stop word lists are often used to allow algorithms to ignore what may be extraneous noise within the corpus of entities to be matched, such as common words which may not add value like "The" or values that are often associated with establishments like "Inc." or "LLC." Preprocessing can aid in stripping out these types of stop words. Alternatively, one may look to algorithmically replace these values, mapping to standardized values, such as "Limited Liability Company" "LLC" "L.L.C." "L L C" standardized as "LLC."

In applying cleansing steps, care needs to be taken in terms of what order different processing steps are taken, along with replacement rules. For example, if stop-word lists result in the omitted word being treated as a space rather than null, it may result in failed matches.

Data Enrichment to Mitigate the Effects of Missing Fields

Toward entity resolution, it may be useful to generate derived attributes to assess data quality, identify potential data quality issues, and to standardize data fields for improved matching. This includes use of geocoding/reverse geocoding engines to generate standardized address fields and latitude/longitude values, along with spatial indexing or other processes for comparing locations to other geographies such as county polygons or jurisdictional boundaries.

Enrichment via geocoding can be a powerful tool to aid in entity matching, however it should be noted that modern geocoding algorithms still suffer from some limitations. As noted in the data elements section, address fields may include non-standard, descriptive types of values, particularly *relative values*, such as "35 miles north of Gunstock on Highway 17" or "Across from the Empire State Building." A geocoding algorithm would need to be sophisticated enough to be able to identify the features and relationships in the descriptive value. While progress continues to be made in this area, most current geocoding algorithms cannot handle these types of relative values. Geocoders can however typically handle references to street intersections.
Geocoders are most adept at handling *absolute values*, such as standard address types. As noted in the data collection section, geocoding engines incorporate algorithms for address parsing, normalization and standardization, for example parsing and standardizing "One South Riverton Avenue" as House Number: "1" Prefix Directional: "S" Street Name: "Riverton" Street Type: "Ave". It should however be noted that geocoders function optimally when input parameters have already been identified as discrete data elements, such as Address / City / State / ZIP. Geocoders also generate match values, returning values corresponding to whether a known match was made at the house number level, street level, street intersection, city or ZIP, providing varying degrees of confidence. Geocoding engines rely on external data sets for matching and comparison, such as Census TIGER data[62] and others, and as such will leverage these data sets to provide a latitude/longitude value, which is typically interpolated along a street segment, with the geocoding data set providing data on address ranges as shown in the figure below:

[62] "Census TIGER data", *US Census*, Retrieved Sep 12 2016
https://www.census.gov/geo/maps-data/data/tiger.html

Figure B7: Street Segment Interpolation

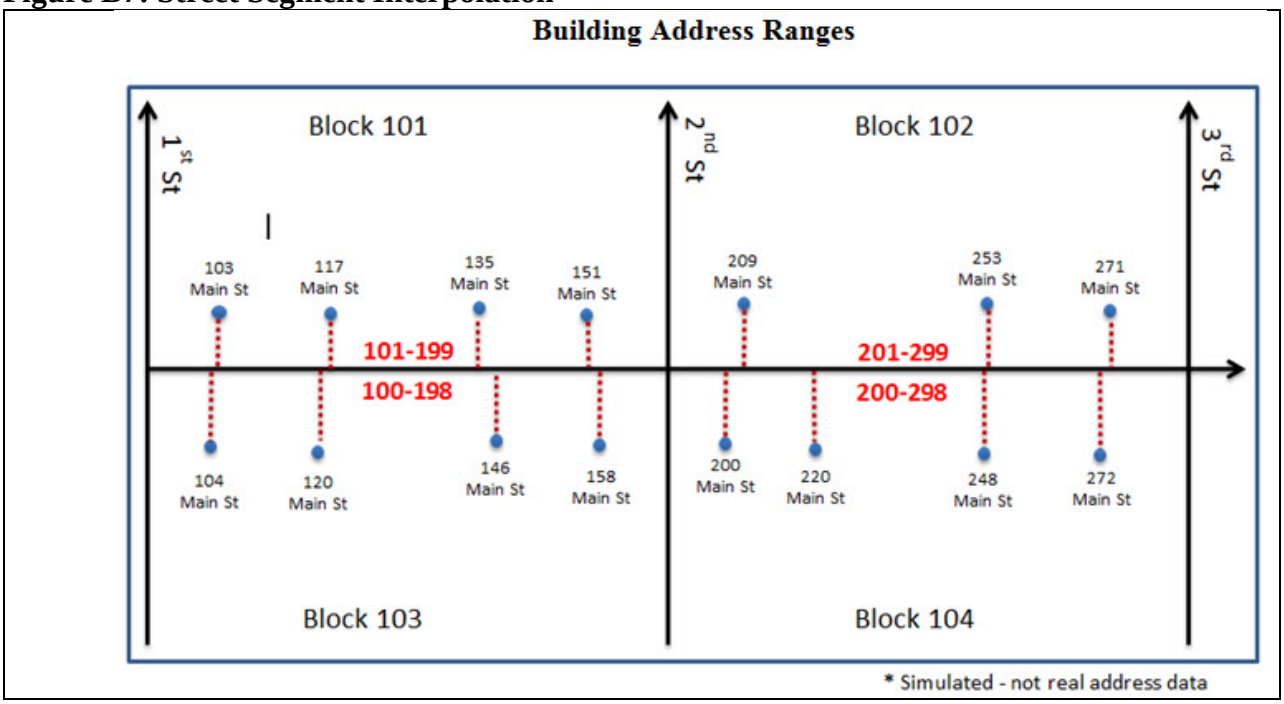


Image source: US Census

Note that the interpolated value calculated will be directly on the street segment centerline. In order to provide a more realistic value, an offset may either be supplied as a default, or may need to be supplied as a parameter to the geocoding engine in order to provide a coordinate offset from the street centerline in the appropriate direction, based on even/odd address number. Some commercial providers such as HERE may also provide point addresses, which attempt to provide actual rooftop or parcel centroid coordinates for addresses. Where these types of match values are available, they will be displayed in the match results.
Geocoding engines can also typically perform *reverse geocoding*, which accepts a latitude/longitude value, and returns the closest matching street address. While caution should be exercised in using at the street address level, this capability can potentially be useful for enriching data which has incomplete or unreliable city/state/ZIP data.

Data to be matched can also be augmented via spatial indexing - this entails testing a location value against polygonal values[63], such as comparing entity latitude/longitude to a county polygon boundary. Spatial databases and GIS tools can provide this capability. Some useful geographies for analysis may include county polygons (note: it is more reliable to use county FIPS codes[64] rather than county names, as county names are

[63] "Point-In-Polygon", *Wikipedia*, Retrieved Sep 12 2016, https://en.wikipedia.org/wiki/Point_in_polygon

[64] "2010 FIPS Codes for Counties and County Equivalent Entities", *US Census Bureau*, Retrieved Sep 12 2016 https://www.census.gov/geo/reference/codes/cou.html

not unique) as well as other Census geographies such as Place[65] or Block[66] (as appropriate, and if other data elements reference these identifiers).

ALGORITHMIC MATCHING METHODS

Deterministic Methods:

The deterministic method is a process of data linkage that requires two records to agree on a pre-determined set of variables to conclude the pair as a link. A match is defined as records from two files that are truly the same unit. A link is defined as two records that are designated as the same unit based on their characteristics and decision rules. While matching on employer name or EIN may seem simple, great caution should be taken to ensure that the data elements from the matched sources are truly comparable and the resulting comparisons are meaningful.

The variables to be used are usually established by subject matter experts and require a significant amount of human review. Deterministic matching works best when there are unique IDs to match, or when there are reliable rules and high quality data elements that can be used for matching, such as email addresses, telephone numbers, dates of birth, NAICS or SIC codes - ideally with multiple elements used in combination - for example, a validated ZIP code by itself might not be sufficient for deterministic matching and would need to be combined with other data elements. It would also potentially be useful in deterministic processing to disambiguate from other similarly-named entities at different locations. Similarly, one could look at relationships within values such as related industrial classifications within NAICS or NAICS-SIC crosswalks[67].

In using deterministic approaches, consider business rules that may affect relationships between entities (for example, a common email or phone number that is at the HQ office but is shared across multiple entities, or one unique email address that shows up across 4,000 chain pharmacy locations).

Relationships between each individual establishment to the HQ office can be established but do not conflate individual entities. Also, as noted in the data elements discussion, consider that there may be third parties involved in the data submission process, such as consultants who are doing preparation on behalf of an establishment, as such, roles, wherever available should be considered. For example, the same law firm may prepare filings for many small companies and may be listed as the “contact person” and “contact mailing address” for all these employers. Linking on this field will create many erroneous links. The usefulness of company websites and Wikipedia pages in deterministic matching of parent companies with subsidiaries should not be understated.

In deterministic matching, it is much more efficient to have automated matching methods err on the side of matching too much, and having an analyst remove erroneous matches than vice-versa. It is much less time-consuming for an analyst to remove erroneous matches than to search out true matches.

---

[65] "Geographic Terms and Concepts - Place," *US Census Bureau*, retrieved Sep 12 2016, https://www.census.gov/geo/reference/gtc/gtc_place.html

[66] "2010 Census - Census Block Maps," US *Census Bureau*, retrieved Sep 12 2016, https://www.census.gov/geo/maps-data/maps/block/2010/

[67] "NAICS to SIC Crosswalk", *NAICS Association*, Retrieved Sep 14 2016 https://www.naics.com/naics-to-sic-crosswalk/

Probabilistic Methods

Probabilistic methods are used to assign a score to two record pairs then set criteria determining link status. Many methods use three ranges of scores, a link with a high score, a non-link with a low score, and a possible link if the score falls in an indeterminate range. The possible links are then reviewed to determine match status.

In attempting to match across large data sets, blocking is an additional strategy that can be used when doing probabilistic record linkage in order to reduce the size of the comparison space. This strategy is used to group similar records before performing the comparison. For example, a full Cartesian join of two files with 450,000 records would yield over 200 billion comparison pairs. It can be seen as these data sets grow the number of comparison pairs can be computationally prohibitive or processing would take an unfeasible amount of time.

There are other strategies that can be used for the probabilistic matching such as standardizing names or fields prior to matching. In some cases edit distance measures can be used to rescale the component weights on a given field instead of giving a field a binary agree/disagree designation. Considerations also need to be made on how to handle missing data when using these methods. In some cases two missing fields should be considered a match while in others they should be non-deterministic or even a non-match. Other ways have been suggested to refine the *m*- and *u*-probabilities described previously using a frequency scaling such as the one described in *Matching and Record Linkage* by William Winkler[68].

Considerations and Clarifications in Probabilistic Entity Resolution Algorithms

1. **Entity Resolution** is not **Classification.** Simply identifying whether or not two records match or do not match is classification while entity resolution develops a dynamic entity using metadata.
2. **Entity Resolution** is not **Clustering.** The goal of clustering is to identify similar groupings of entities; the goal of entity resolution is to reconcile different iterations of the same entity down to their common iteration.

Algorithms for Comparing Latitude/Longitude

Comparisons of latitude/longitude pairs for proximity assume that coordinates contain valid values. Wherever possible, a prior data QA step should be taken to ensure that values pass reasonable checks. Common issues specific to entered lat/long values include reversal of lat/long values to long/lat, omission of sign for hemisphere, and use of placeholder values, such as 0,0 or 1,1. Additionally, data sets may use differing standards, such as sexagesimal degrees-minutes-seconds versus decimal degrees. These will need to be parsed and converted to decimal values. One way to perform a validation is to use a reverse geocoder on the address and compare the reverse geocoding result to the provided latitude/longitude value.
Haversine Distance: The Haversine Distance formula[69] can be used to compare spatial latitude/longitude tuples by computing Great Circle surface distance[70] between locations.

---

[68] Winkler, W. (2000). *Frequency-Based Matching in Fellegi-Sunter Model of Record Linkage*

[69] Williams, Ed (2011), "Aviation Formulary V1.46," Accessed Sep 12 2016
http://williams.best.vwh.net/avform.htm#Dist

[70] "Great Circle," *Wikipedia*, Accessed Sep 12 2016
https://en.wikipedia.org/wiki/Great-circle_distance

$$d = 2\sin^{-1}\left(\sqrt{\sin\left(\frac{\varphi_1 - \varphi_2}{2}\right)^2 + \cos(\varphi_1)\cos(\varphi_2)\sin\left(\frac{\lambda_1 - \lambda_2}{2}\right)^2}\right)$$

where:

- φ_1 is first coordinate latitude in radians, φ_2 is second coordinate latitude in radians
- λ_1 is first coordinate longitude in radians, λ_2 is second coordinate longitude in radians
- d is computed angular distance between the points

Text Matching Algorithms for Entity Resolution

The following methods address challenges related to missing important data fields, or having inconsistent data formats in text fields, such as name and address fields. Depending on which data element is being matched across data sets, different algorithms can be applied.

Objective	Description	Examples
String Matching	Quantify permutations needed to convert one string to another	Edit Distance, Alignment, Phonetic
Distance Metrics	Apply physical distance measures to abstract concept of data objects	Similarity, Text Analytics
Relational Matching	Conjunctive view reliant on one data object's relationship to other objects	Set Based, Aggregate

String Matching

String matching algorithms are concerned with whether or not two strings say the same thing. Outlined in the table below, there are four essential approaches. They may, however, be further subdivided into exact element-by-element character or phonetic comparison.

Boolean Matching, is easily understood as a Yes or No, 0 or 1, match or non-match between two strings. It is the most simplistic of the group and is the core logic on which the subsequent algorithms operate.

Method	Description	Examples
Edit Distance	Quantified permutations to convert one textual string into another	Levenshtein, Jaro-Winkler
Jaccard Coefficient	Ratio of existence or absence of one entity's individual attributes in another	Jaccard
Phonetic Similarity	Pronunciation of letters are phonetically related on a 0 to 1 similarity scale, aka, fuzzy matching	Soundex, Translation

In the heavily-studied **Edit Distance**, similarity is quantified by physically measuring the permutations needed to convert one string into other. The core implementation of edit distance is Levenshtein, which penalizes for insertions, deletions and substitutions. Over time Levenshtein has been modified with additional costs for gaps (Sellers), transpositions (Smith-Waterman), and affine gaps, i.e., weighted costs per each of the actions or the location of where the permutation must be made (Gotoh).

Jaro Distance is a hybrid version of edit distance whose more popular counterpart, Jaro-Winkler, is considered a hybrid algorithm. Practically, Jaro slides along two strings, comparing nGrams along the way to quantify the number of characters appearing in the same position and the number of transpositions required for coincident

characters which must be reordered in one string to match the other. Its best application is with short strings, also it disobeys triangle inequality.
The **Jaccard Coefficient** is an element-by-element measure of intersection. Stated otherwise, it is the ratio of the intersecting set to the union set. The Jaccard Coefficient satisfies triangle inequality. One frequently-confused issue: the similarity version of Jaccard and Tanimoto Coefficients are identical, but their dissimilarity coefficients diverge due to triangle inequality. While this justifies the need for two separate algorithms, they are frequently credited as the Jaccard-Tanimoto Coefficient as both mathematicians independently published this ratio unbeknownst of each other.
Phonetic Similarity algorithms result in Soundex encodings, which sidestep misspellings and variations, by indexing a table of language-specific homophones for a string's Soundex encoding rather than searching the string itself. Two critical inputs to phonetic similarity are (1) discerning which language the string is written in and (2) knowing the context of the letters you are matching. The crucial former prerequisite is accomplished by matching pronunciation rules of letter sequences using their location in the string ("sch" in German vs. "sz" in Polish at beginning of a string). The latter is accomplished by parsing the string into a sequence of phonetic tokens according to pronunciation rules in that language. The International Phonetic Alphabet (IPA) is popularly used to identify tokens with corresponding sounds, though frequently criticized for being too fine of match.

Distance Metrics

While string matching compares strings element-wise, distance metrics incorporate a spatial element, measuring the literal distance between two entities using algorithms seen in the table below. The first three are inter-related, easy visualized by plotting the entities to be reconciled on a preference space with x and y axes. A discerning eye anticipates the obvious limitation of these, that only a certain number of attributes is practical.

Method	Description	Examples
Euclidean	'As-the-crow-flies' distance	L_2-Norm, Ruler, Spearman
Mahalanobis	Matching for centered and standardized distances	
Manhattan	Distance if following a grid-like path, turning corners	L_1-Norm, Taxicab, City-Block, Footruler, Rectilinear
Minkowski	Generic edit distance, of which Euclidean, Manhattan and Chebyshev are instances	Soundex, Translation
Chebyshev	Distance along axis on which the objects show greatest absolute difference	L_{max}-Norm, Chessboard
Text Analytics		Pearson Coefficient, Jaccard Similarity Coefficient
Vector Similarity		Cosine Similarity, TFIDF

Minkowski Distance[71] is the generalized distance between two points in a plane. Specialized forms include Euclidean, Manhattan and the less-common Chebyshev.
Mathematically, ***Euclidean Distance*** is Minkowski Distance squared. Practically, it is the equivalent of the bishop in chess in that it moves diagonally, or as-the-crow-flies. The Euclidian Squared Distance Metric is a variation with quicker processing time since it does not take the square root.

[71] "Minkowski Space," *Wikipedia*, accessed Sep 12 2016 https://en.wikipedia.org/wiki/Minkowski_space

Manhattan Distance is mathematically the Minkowski Distance raised to 1; it is the same as Euclidean, except for the requirement of absolute value since it is not squared. Practically it is the equivalent of a knight, which makes L-shape moves. Its name is coined after the great borough of New York City, where pedestrians and cars must obey the laws of street corners.

Text Analytics

In contrast to the Minkowski distances, which scale similarity on a scale of 0 to 1, Pearson's Coefficient scales from –1 to 1, in other words fitting similarity along a line, making it a better choice for non-normalized data and when attributes' scales are undefined. Mathematically, it is the ratio between two points' covariance and standard deviation.
The ***Jaccard Similarity Coefficient*** is mathematically the size, i.e., the existence of defined attributes using a binary 0/1, of the intersection of two points divided by the size of the union of the points.

Vector Similarity

First, a quick introduction to ***Vector Similarity***. We construct a VSM (Vector Space Model) as a series of vectors quantifying frequency of a selected attribute inside a document. These vectors are subsequently assembled into a matrix, allowing easy algebraic manipulation. Two vector similarity functions are of particular note:
The widely-known bag of words model is enhanced to a 'bag of terms' with ***TF-IDF***. Weighted TF-IDF incorporates local and global parameters, applying a logarithmic scale to account for a term's relative importance versus frequency of appearance. This allows the algorithm emphasize less-frequent terms' importance. TF-IDF normalizes any bias introduced into the vectors by keyword spanning, most commonly with the L2 (Euclidean) Norm. The equation is the row-wise multiple of two matrices: TF (Term Frequency, the local parameter): matrix of vectors of selected terms' frequencies of appearance in each document IDF (Inverse Document Frequency, the global parameter): diagonal matrix version of vector containing, for each term, the log of the number of documents divided by the number of documents in which the selected term appears
Cosine Similarity is most useful when it is known that two points have a high proportion of non-shared attributes. Mathematically, the attributes are presented in a vector, allowing the algorithm to find the dot product of the two points. It measures the angle of the vector rather than the magnitude. Theoretically this results in the angle between the two points' attributes; a 90° angle is perfect dissimilarity.

Relational Matching

Relational Matching algorithms retain many commonalities with Jaccard and Euclidean, but are mathematically differentiated since as a group they do not satisfy triangular inequality. Practically speaking, while the aforementioned algorithms measure similarity between two documents, relational algorithms broaden the playing field, incorporating a third document's attributes into the mix.
While the ***Tanimoto (Jaccard) Similarity Coefficient*** is the same as the Jaccard Similarity Coefficient, the dissimilarity coefficient is where these two algorithms diverge. This is to say that Tanimoto is a proper similarity metric but its distance metric is not mathematically legal since it allows the two points to share commonality with a third point, causing it to disprove triangular inequality. In application, Tanimoto is preferred over Jaccard in cases when we want to allow the two points, themselves very different, to share commonalities with a third point. Mathematically, Tanimoto is the number of intersecting elements divided by the number of elements in either point.
Dice's Coefficient is mathematically the number of intersecting attributes divided into the total population of attributes, thus, as with Tanimoto, it shares a definition in its similarity metric version but Dice's dissimilarity

coefficient is not the same as it does not satisfy triangle equality. Compared to Markowski's, Dice's coefficient is sensitive to heterogeneity in data sets and less sensitive to outliers.
A simplistic similarity measure is ***Common Neighbors***, which predicts the likeness between two documents in terms of the number of common attributes each of those two documents independently shares with other documents.
Adamic/Adar Weighted modifies Common Neighbors to weight attributes that are shared infrequently relatively higher than those which are more common across all documents. Mathematically, this is accomplished by weighting a shared attribute's vector value with 1/ log of the number of times the attribute is shared across all documents.

Hybrid Metrics

Experts inevitably merge foundational seminal concepts together. Thus, this section explains some well executed hybrid metrics derived from the ones above.

Method	Description
Jaro-Winkler	Jaro Distance modified to favor common prefixes
Monge-Elkan	Atomic Strings matching with Gotoh
Soft-TFIDF	A forgiving version of Cosine & Monge-Elkan

Jaro-Winkler is a hybrid algorithm with its roots in Jaro Distance, edit distance, but incorporates Cosine Similarity's approach towards strings with high degree of dissimilarity and TF-IDF's concept of applying a weight to certain elements. It improves on the basic Jaro Distance by accommodating for strings with a common prefix, effectively biasing its matching to favor similarity between two otherwise-dissimilar strings who share a common prefix.
Monge-Elkan is sometimes considered synonymously with Smith-Waterman Edit Distance, but the two are differentiated as Monge-Elkan uses the Gotoh Distance. The confusion is understandable, as Gotoh amends Smith-Waterman distance by accommodating affine gaps. Practically, it applies the combined power of Levenshtein and Jaro Similarity Measures to n-Gram subsets of strings (called atomic strings). Mathematically, Monge-Elkan uses Gotoh edit distance to evaluate atomic strings against each other. Before deciding on Monge-Elkan you should understand how sensitive your matching is to the symmetry of your strings, i.e., if one string is longer than the other. It has quadratic time complexity due to its recursive calculations.
Soft-TFIDF adds a forgiveness factor to Cosine Similarity and Monge-Elkan, which are intolerant of spelling errors as they roll along atomic strings in the order of appearance by incorporating TF-IDF's concept of matrix of terms (i.e., letters) to develop an internal frequency per Atomic string. Soft-TFIDF calculates an inner score comparator, thus allowing partial matches.

Fellegi-Sunter Method and the Expectation-Maximization (EM) Algorithm

The Fellegi-Sunter is a common and well established method of probabilistic record linkage. This method is used to develop the scores for determining link status. The Fellegi-Sunter model sums the weights the log likelihood for each component to determine a match score. The log likelihoods are developed by taking the log of the ratio of the m- and u-probabilities. The m- probability is the probability of agreement on a field between two records that are a true match. The u-probability is the probability of agreement on a field between two records that are not a true match. The single component weight is $\log\left(\frac{m}{u}\right)$, if the field agrees and $\log\left(\frac{1-m}{1-u}\right)$, otherwise. The fields with more distinguishing power will have lower u-probabilities and therefore yield a large weight if the fields agree. There are a few different ways to estimate these probabilities. The u

probabilities can be estimated as a ratio of the frequency of the values divided by the number of pairs in the comparison space. The m probabilities can be estimated by taking samples of pairs and calculating a match rate based on human review or can be based on prior knowledge. A common method would be to use the EM algorithm to estimate both the m- and u-probabilities using the observed agreement patterns in the data. The EM algorithm starts with a sometimes arbitrary estimate of the m- and u-probabilities along with an estimate of the true match rate. The first step or expectation step is to use these initial parameters to estimate the probability of observing an agreement pattern among all the components given they are a true match for each record pair. Using these probabilities and the estimated match rate, the probability of a true match given the observed agreement pattern is calculated. Next the complete log-likelihood is separated into three maximization problems to solve for the new estimates of the m- and u- probabilities along with the true match rate. This process is repeated until some convergence criteria is met. See *Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage*[72] by William Winkler to read this process is detail.
Naive Bayes Machine Learning: Naive Bayes Machine Learning methods have also been explored for entity resolution[73].

Ensemble Methods

In addition to many of the individual techniques described above, ensemble or composite approaches may help to enhance the quality of matching efforts. Given that many of the matching tasks are on text fields or text within fields, and string distance measures are often used to quantify the degrees of similarity and dissimilarity, leveraging the benefit of several measures may yield better results than relying on any one measure. Previous work (see Tejada et. al. 2001, Cohen and Richman 2002, and Bilenko and Mooney, 2002 in the supplemental bibliography) has empirically shown that compositing individual measures may offer better performance. A simple example of this method is provided with the code examples, and represents one of many possible approaches that could be taken in looking at finding establishments in data with a high degree of variability in how the establishment names and addresses are coded.

Evaluating Results of Matching and Entity Resolution Approaches

Cutoff Scores

When linking data, there could be an unfeasible amount of records that could be flagged for review given the size of the data sets being linked. Given consideration to the amount of resources that would be needed to review the links there are additional methods that reduce the amount of human review. Dusetzina, Tyree, Meyer, et al.[74] suggests a single cutoff developed by Cook[75] that uses a single cutoff which allows an acceptable distance between the starting weight and the desired weight. The acceptable distance is determined by the researcher's business need on the desired specificity and sensitivity. To increase the number of true matches at the cost of introducing more false positives the researcher would use a more liberal cutoff. If the

[72] Winkler, W. (2002). *Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage*

[73] Yun Zhou , Minlue Wang, Valeriia Haberland, John Howroyd, Sebastian Danicic, and J. Mark (2015) *Improving Record Linkage Accuracy with Hierarchical Feature Level Information and Parsed Data*
https://research.gold.ac.uk/17342/1/Yun_AMBN_2015_journal.pdf

[74] Dusetzina, SB., Tyree, S., Meyer, AM., et al. (2014). *Linking Data for Health Services Research: A framework and Instructional Guide*

[75] Cook LJ, Olson LM, Dean JM. (2001) *Probabilistic record linkage: relationships between file sizes, identifiers, and match weights.* Methods Information Med. 2001;40:196-203. PMID: 11501632.
http://www.ncbi.nlm.nih.gov/books/NBK253313/pdf/Bookshelf_NBK253313.pdf

goal is to only keep record pairs with high probabilities of a match the researcher would set a more conservative cutoff at the risk of increasing false negatives.

Other considerations for cutoffs must be made on whether the matches should be one-to-one, one-to-many, or many-to-many. In the case of one-to-one matches the “greedy” strategy can be employed. The greedy strategy accepts the best match therefore it can only be used for one-to-one matching. This involves taking the pair that has the highest score for a given record. This can be employed without a cutoff if the assumption is a match exists for every record in a given file. In addition a combination of using the cutoff with the greedy method can be used when the assumption is a match may or may not exist and if it does there will only be one.

Performance Evaluation

A common problem with the probabilistic method is a way to evaluate the quality of the links. In most cases the sensitivity, specificity, positive predictive value and negative predictive value are analyzed. There are different ways to calculate these such as taking samples of the pairs and manual reviewing the pairs and characterizing them as true/false positive and true/false negative. However, this process is very resource intensive. Another option would be to use a training set to evaluate the method using known matches and non-matches but this scenario is usually not available.

Machine Learning in Evaluation

Another evaluation option would be to use a training set to evaluate the method using known matches and non-matches but this scenario is usually not available, however if a curated data set already exists, this may be an option. A machine learning exercise can help in evaluating weighting and thresholds to be applied in tuning some of the other matching strategies, where a data set containing verified matches would be used to train the algorithm and then be used to analyze values returned by the match algorithms to aid in determining what combinations of algorithms are effective; what thresholds and tunings should be used for each algorithm; and weightings and approaches, whether hierarchical or otherwise should be used in using multiple algorithms in combination.

APPENDIX C: Best Practices - Bibliography

Bengfort, B. (2008). "Entity Resolution for Big Data: A Summary of the KDD 2013 Tutorial Taught by Dr. Lise Getoor and Dr. Ashwin Machanavajjhala." http://www.datacommunitydc.org/blog/2013/08/entity-resolution-for-big-data

Bengfort, B. (2015). "A Primer on Entity Resolution." http://www.slideshare.net/BenjaminBengfort/a-primer-on-entity-resolution

Bilenko, M and Mooney, R. (2003). "Adaptive duplicate detection using learning string similarity measures." Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Bilenko, M. (2006). "Learnable Similarity Functions and Their Application to Record Linkage and Clustering." http://www.cs.utexas.edu/~ml/papers/marlin-dissertation-06.pdf

Bilgic, M., Licamele, L., Getoor, L. and Shneiderman, B. (2006). D-Dupe: An interactive tool for entity resolution in social networks. Proceedings of IEEE Symposium on Visual Analytics Science and Technology.

BYU Data Mining Lab - Record Linkage Resources
https://facwiki.cs.byu.edu/DML/index.php/Record_Linkage_Resources

Chen, Z., Kalashnikov, D., and Mehrotra, S. (2009). "Exploiting context analysis for combining multiple entity resolution systems." Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data.

Christen, P. (2008). "Automatic record linkage using seeded nearest neighbour and support vector machine classification." Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Cochinwala, M. (2001). "Efficient data reconciliation." Information Sciences 137.1, 1-15.

Cohen, William W., and Jacob Richman (2002). "Learning to Match and Cluster Large High-Dimensional Data Sets for Data Integration." In KDD 2002: 475-480. http://www.cs.cmu.edu/~wcohen/postscript/kdd-2002.pdf

Daggy, J., Xu, H., et al. (2013). *A practical approach for incorporating dependence among fields in probabilistic record linkage*

"D-Dupe: A Novel Tool for Interactive Data Deduplication and Integration."
http://linqs.cs.umd.edu/projects/ddupe/

Elmagarmid, Ahmed K., Ipeirotis, P., Verykios, V. (2007) "Duplicate Record Detection: A Survey."
https://www.cs.purdue.edu/homes/ake/pub/TKDE-0240-0605-1.pdf

Grannis, S., Overhage, M., et al. (2003). *Analysis of a Probabilistic Record Linkage Technique without Human Review*

Hadjieleftheriou, M. and Srivastava, D. (2010). “Weighted Set-Based String Similarity.” http://sites.computer.org/debull/A10mar/divesh-paper.pdf

Herzog, T., Scheuren, F. and Winkler, W. (2007). “Data quality and record linkage techniques.” Springer Science & Business Media.

Fellegi, I. and Sunter, A. (1969). "A theory for record linkage." Journal of the American Statistical Association 64:328. http://courses.cs.washington.edu/courses/cse590q/04au/papers/Felligi69.pdf

Gupta, R. and Sarawagi, S. (2009). “Answering table augmentation queries from unstructured lists on the web.” Proceedings of the VLDB Endowment 2.1, 289-300.

Kang, H., Getoor, L. and Singh, L. (2007). C-Group: A visual analytic tool for pairwise analysis of dynamic group membership. Proceedings of IEEE Symposium on Visual Analytics Science and Technology.

Kang, H., Sehgal, V. and Getoor, L. (2007). GeoDDupe: A novel interface for interactive entity resolution in geospatial data. Proceedings of Information Visualisation, 489-496.

Kang, H., Getoor, L., Shneiderman, B., Bilgic, M. and Licamele, L. (2008). “Interactive entity resolution in relational data: A visual analytic tool and its evaluation.” IEEE Transactions on Visualization and Computer Graphics, Volume 14, Number 5, 999-1014.

Kardes, H., Konidena, D., Agrawal, S., Huff, M., and Sun, A. (2013). “Graph-based Approaches for Organization Entity Resolution in MapReduce.” http://www.aclweb.org/anthology/W13-5010

Navarro, G., Baeza-Yates, R., Sutineny, E., and Tarhioz, J. (2001). “Indexing Methods for Approximate String Matching.” http://www.dcc.uchile.cl/~gnavarro/ps/deb01.pdf

Raffoa, J. and Lhuillery, S. (2009). “How to Play the ‘Names Game’: Patent Retrieval Comparing Different Heuristics.” http://infoscience.epfl.ch/record/161961/files/Lhuillery%20RP2009b.pdf

Ravikumar, P. and Cohen, W. (2004). “A hierarchical graphical model for record linkage.” Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. AUAI Press.

“Record Linkages.” https://en.wikipedia.org/wiki/Record_linkage

Sarawagi, S and Bjamidipaty. (2002). “Interactive deduplication using active learning.” Proceedings of the Either ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Steorts, R., Hall, R., Fienberg, S. (2015). "A Bayesian Approach to Graphical Record Linkage and De-duplication." http://arxiv.org/abs/1312.4645

Tejada, S., Knoblock, C., and Minton, S. (2001). “Learning object identification rules for information integration.” Information Systems 26.8, 607-633.

Winkler, W. (2006). “Overview of record linkage and current research directions.” Bureau of the Census.

Winkler, W. and Thibaudeau, Y. (1990). “An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Decennial Census.” https://www.census.gov/srd/papers/pdf/rr91-9.pdf

Winkler, W. (2014). *Matching and Record Linkage*
Zhu, V., Overhage, M., et al. (2009). *An Empiric Modification to the Probabilistic Record Linkage Algorithm Using Frequency-Based Weight Scaling*

APPENDIX D: Workgroup Methodology

Drawing on strong interest among Federal agencies to achieve efficiencies in matching U.S. employer data across Federal data sets for data analysis, evaluations, and statistical activities, based on common needs across agencies, OMB convened the Employer Data Matching Workgroup in 2016 to complete the following tasks:

- Document pain-points related to matching and uniquely identifying establishments and firms within and between data sets and over time. For example, how are agencies classifying employers in their data and what terms are interchangeable across data sets (e.g., establishment level firm, or enterprise)? What challenges exist in the data for creating matches?
- Identify current successful strategies used by agencies to address this challenge in the context of analyzing data, conducting evaluations, producing statistics, and identifying where additional strategies may be needed to further facilitate this work. This process focuses on coming to agreement on one or a few methods that will be effective for multiple agencies.
- Along with OMB staff, work to develop a white paper identifying best practices, and high-level implementation steps, on how Federal agencies can achieve efficiencies in identifying and matching unique firms and establishments (and the relationship between the two) within and across Federal data sets, for the purposes of analyzing data, conducting evaluations, and producing statistics.

The Workgroup has representative and cross-functional mixture of:

- statistical agencies reporting data on employers,
- evaluation offices examining employers, and
- agencies with Federal programs affecting employers whose data are prime for research, evaluation, and data analysis.

See Table D1 for a complete list of participating offices and component agencies. OMB has also provided strong support for this work. The Social Security Administration’s Office of Data Exchange also provided subject matter expertise to help inform the Workgroup's development of best practices.

Table D1: Employer Data Matching Workgroup, Participating Offices and Component Agencies

Department/Agency	Component Agency or Office
Department of Agriculture	Economic Research Service
Department of Commerce	Bureau of Economic Analysis
Department of Commerce	Census Bureau
Department of Commerce	Commerce Data Service
Department of Commerce	International Trade Administration
Department of Commerce	Minority Business Development Agency
Department of Commerce	National Oceanic and Atmospheric Administration
Department of Commerce	Office of the Secretary, Office of Performance, Evaluation, and Risk Management
Department of Education	Institute of Education Sciences, National Center for Education Statistics
Department of Health and Human Services	Agency for Healthcare Research and Quality
Department of Health and Human Services	Centers for Medicare and Medicaid Services, Center for Medicare & Medicaid Innovation
Department of Housing and Urban Development	Office of Policy Development and Research
Department of Labor	Bureau of Labor Statistics
Department of Labor	Chief Evaluation Office
Department of Labor	Employee Benefits Security Administration
Department of Labor	Occupational Safety and Health Administration
Department of Labor	Wage and Hour Division
Department of the Treasury	Internal Revenue Service, Statistics of Income Division
Department of the Treasury	Office of Financial Research
Department of the Treasury	Office of Tax Analysis
Department of Transportation	Chief Data Officer
Environmental Protection Agency	National Center for Environmental Economics
Environmental Protection Agency	Office of Environmental Information
Equal Employment Opportunity Commission	Office of Information Technology
Equal Employment Opportunity Commission	Program Research and Surveys Division
General Services Administration	Federal Acquisition Service
Securities and Exchange Commission	Division of Economic and Risk Analysis
Small Business Administration	Office of Performance Management

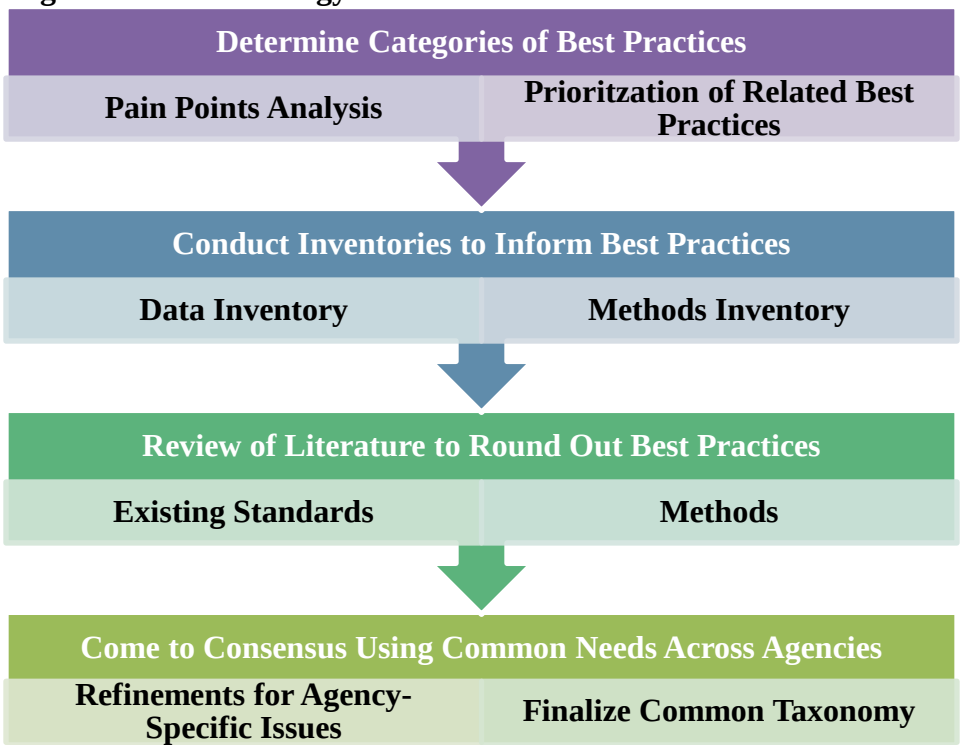
Figure D1 provides an overview of the Workgroup's approach. The Workgroup first prioritized agency pain points, and potential categories of best practices, to determine where to focus its efforts. Through this work, it became clear that there was consensus to review best practices for matching, and to review of long-term, high-value changes that agencies could implement to improve matching.

Given the unique scoping and makeup of the Workgroup, members recommended conducting a Workgroup-specific Data Inventory to get information on representative data sets (at a more detailed level than information provided through preexisting efforts, such as Data.gov), and a Workgroup-specific Methods Inventory to obtain information on representative methods for record linkage and entity resolution, to inform best practices.

- The Workgroup relied primarily on the Data Inventory to generate sections in the white paper related to long-term best practices. In August and September of 2016, Workgroup members provided information on a representative sample of data sets from their agencies contain information on individual employers, firms, and/or establishments that have the widest coverage, greatest use, or which they were most interested in matching to other data sets.
- The Workgroup relied primarily on its Methods Inventory to generate Best Practices for Matching and Data Collection Improvements. The Methods Inventory collected and disseminated best practices for matching. Specifically, the Workgroup asked for information on successful methods and tools agencies use for achieving efficiencies in matching employer, establishment or firm data.

The Workgroup then supplemented synthesized findings with iterative feedback from member agencies, a literature review, and an interagency clearance process. This approach was deemed analytically sufficient by the Workgroup's co-chairs, OMB, and Workgroup members.

Figure D1: Methodology



The Data Inventory produced information on administrative and statistical data sources from a subset of participating Agencies, as shown in Table D2:

Table D2: Agencies and Offices Represented in Data Inventory

Department/Agency of Data Source	Component Agency or Office of Data Source
Department of Commerce	Bureau of Economic Analysis, Balance of Payments Division
Department of Commerce	Bureau of Economic Analysis, Direct Investment Division
Department of Commerce	Census Bureau
Department of Commerce	Census Bureau (Survey sponsored by the Agency for Healthcare Research and Quality)
Department of Education	Institute of Education Sciences, National Center for Education Statistics
Department of Education	Office for Civil Rights and National Center for Education Statistics
Department of Health and Human Services	Administration for Children and Families, Office of Child Support Enforcement
Department of Labor	Bureau of Labor Statistics, Office of Employment and Unemployment Statistics
Department of Labor	Employee Benefits Security Administration
Department of Labor	Occupational Safety & Health Administration
Department of Labor	Wage and Hour Division
Department of Transportation	Federal Aviation Administration
Department of Transportation	Federal Motor Carrier Safety Administration
Department of Transportation	Federal Railroad Administration
Department of Transportation	National Highway Traffic Safety Administration
Department of Transportation	Pipeline and Hazardous Materials Administration, Office of Hazmat Safety
Department of Treasury	Statistics of Income Division, Internal Revenue Service
Equal Employment Opportunity Commission	Program Research and Surveys Division, Office of Research, Information and Planning
General Services Administration	Integrated Award Environment
Securities and Exchange Commission	(Multiple Offices)
Small Business Administration	Office of Capital Access
Small Business Administration	Office of Disaster Assistance
Small Business Administration	Office of Entrepreneurial Development

72% of the data sources were administrative data sources and not subject to CIPSEA's requirements, but are, or could be, confidential under other statutes or policies; 28% of the data sources were from statistical agencies and subject to CIPSEA's requirements. Among the administrative data sources, the main unit of analysis is at the transactional level, and employers or firms referenced in these sources can be at either the establishment or

enterprise level. Among the statistical data sources, six (55%) were identified as at, or generally at, the establishment level; three (27%) were at the enterprise level; the remaining sources were at another type of level (e.g. government unit). The Workgroup examined commonalities among fields available in the example sources flagged from these agencies to develop common data elements. For example, approximately one-half of the sources in the inventory noted at least one identifier which could be used across agencies (e.g. EIN). Approximately one-half of the sources had an agency-specific identifier. As about half of the example sources included at least one interagency identifier, the Workgroup proceeded to include it in the Common Data Elements. 87% of the sources included physical and/or mailing addresses, and nearly all the sources included name fields.

As further described in Appendix A, the Workgroup also reviewed the Data Inventory for suggestions on potential authoritative sources, and considerations for reviewing legal barriers, policy and legal interpretations, and capacity and resource constraints.

To gather sufficient information to describe best practices in employer record linkage and entity resolution, the Workgroup conducted a methods inventory: a questionnaire that was sent to a broad spectrum of Federal agencies and offices in July 2016. Responses were obtained from 21 individuals representing 14 different Federal agencies or offices. The results are further detailed in Appendix B.

In order to round out the findings from the Data and Methods Inventories, the Workgroup also made note of relevant sources, and reviewed them in compiling best practices. These sources are listed in Appendix C.

Future work would include further refinement of best practices. Member agencies agreed that the approaches we capture are useful and agencies stand to gain from implementing them. There is however, an open question on how best to achieve the benefits of the best practices. A second phase of the work would focus on launching a methods community for matching employer data to develop refined methods by application, and further vetting of long-term best practices for consideration in future policy documents.

Appendix E: Best Practices Code Examples

Agencies submitted these code examples to support a better understanding of current methods and best practices, and illustrate approaches used in data remediation, canonicalization and data matching. This appendix provides the code samples in full so that individuals seeking to match data on employers can see the specific steps that are taken. Specifically:

- Code Example 1: Census SAS code to canonicalize/standardize string content to facilitate matching
- Code Example 2: Code from Census in SAS to demonstrate matching with Business Register (This code requires the SAS Data Quality Server)
- Code Example 3: Code from CEO/DOL to implement probabilistic matching, EM algorithm (This code does NOT require the SAS Data Quality Server)
- Code Example 4: Stata code from EBSA/DOL to remediate data quality issues and match data, using regular expressions
- Code Example 5: Code from OSHA/DOL to normalize/canonicalize/remediate data quality issues. This uses regular expressions, and makes use of information from some fields that code example 1 deletes.
- Code Example 6: Code from CEO/DOL to composite string distance measures
- Code Example 7: Code compiled by Rebecca Bilbro, in Python, to test string matching functions
- Code Example 8: Module used by the Office of Foreign Labor Certification (OFLC), ETA/DOL to remediate data quality issues

Note that many of the code examples that deal with data quality issues search for very specific patterns which can cause issues in both exact matching and inexact matching applications, as opposed to searching for generally unexpected or anomalous values. The exact code syntax is less important, rather the examples that warrant modification illustrate the kinds of actual agency data issues and problems that impede or limit the effectiveness of matching. The goal of these code sections is typically to standardize strings for comparison so that exact matching can occur, or to increase the accuracy and precision for inexact matching processes by reducing the rates of false positives and false negatives.

Code example 1: Census SAS code to canonicalize/standardize string content to facilitate matching

```
/* Cleaning up the NETS - Standardizing the NETS */

options compress=yes;
libname NETS '; *where the data is located;
run;
libname output ';
run;  * output files libname;

%let state=CA; *CHOOSE THE STATE

/*///// RAW NETS DATA /////*/

proc sort data=nets.nets_&state.2007; * RAW DC/CA FULL DATA FROM 1990 TO 2007 ;
by Company;
run;

/*///// NETS ONLY HQ /////*/
```

```
* Keeping only a subset (headquarters) of the observations;
data nets_&state.2007hq;
set nets.nets_&state.2007(where=(category='Standalone' or category='Headquarters')) ;
run;

/* Only keeping some NETS variables */
data Nets_&state.2007no (keep=Company Address City State ZipCode);
        set Nets_&state.2007hq(drop= xxx);
run;

proc sql;
        create table Nets_&state.2007nodups as
        select distinct *
        from Nets_&state.2007no;
quit;

/* Keeping only firms in STATE */
data Nets_&state.2007nodups;
        set Nets_&state.2007nodups;
        if state^="&state." then delete;
run;

 /*////////////// Further Cleaning and Pairing ////////////////*/
/* IF THE DATA IS NOT CAPITALIZED WE NEED TO CAPITALIZED IT BEFORE */

%DQLOAD(DQLOCALE=(ENUSA), DQSETUPLOC="sas pathname");

%macro standard(n,s,file);
data &file._st;
        set &file;
        &n=tranwrd(&n,".","");
        &n=tranwrd(&n,",","");
        &n=tranwrd(&n,"& ","");
        &n=tranwrd(&n,"&","");
        &n=tranwrd(&n,"'","");
        &n=tranwrd(&n,"#","");
        &n=tranwrd(&n,"-","");
        &n=tranwrd(&n,"/","");
        &n=tranwrd(&n,"  "," ");
      &n=tranwrd(&n,' INC ', '');
      &n=tranwrd(&n,'LLC ', '');
        &s=tranwrd(&s,".","");
```

```
    &s=tranwrd(&s,",","");
    &s=tranwrd(&s,"& ","");
    &s=tranwrd(&s,"&","");
    &s=tranwrd(&s,"'","");
    &s=tranwrd(&s,"#","");
    &s=tranwrd(&s,"-","");
    &s=tranwrd(&s,"/","");
    &s=tranwrd(&s,"  "," ");
    &s=tranwrd(&s,"P O ", "PO ");
    &n=tranwrd(&n,' THE ',");
&n=tranwrd(&n,'THE ',");
&n=tranwrd(&n,' OF ',");
&n=tranwrd(&n,","," ");
&n=tranwrd(&n,"'","");
&n=tranwrd(&n,"-"," ");
&n=tranwrd(&n,"&"," ");
&n=tranwrd(&n,' CO ',");
&n=tranwrd(&n,' COMPANY ',");
&n=tranwrd(&n,' CORP ',");
&n=tranwrd(&n,' INC ',");
&n=tranwrd(&n,' L P ',");
&n=tranwrd(&n,' LP ',");
&n=tranwrd(&n,'LLC ',");
&n=tranwrd(&n,'L L C ',");
&n=tranwrd(&n,' PL ',");
&n=tranwrd(&n,' P L ',");
&n=tranwrd(&n,'PLLC ',");
&n=tranwrd(&n,'P L L C ',");
&n=tranwrd(&n,'GRP ',");
&n=tranwrd(&n,'LTD ',");
&n=tranwrd(&n,'LLP ',");
&n=tranwrd(&n,' AND ',");
&n=tranwrd(&n,' ASSOC ',");
&n=tranwrd(&n,' ASSOCS ',");
&n=tranwrd(&n,' DBA ',");
&n=tranwrd(&n,' D B A ',");
&n=tranwrd(&n,' MD ',");
&n=tranwrd(&n,' M D ',");
&n=tranwrd(&n,' DMD ',");
&n=tranwrd(&n,' D M D ',");
&n=tranwrd(&n,' PS ',");
&n=tranwrd(&n,' P S ',");
&n=tranwrd(&n,' MSD ',");
```

```
&n=tranwrd(&n,' M S D ',");
&n=tranwrd(&n,' DVM ',");
&n=tranwrd(&n,' D V M ',");
&n=tranwrd(&n,' ESQ ',");
&n=tranwrd(&n,' DDS ',");
&n=tranwrd(&n,' D D S ',");
&n=tranwrd(&n,' DR ',");
&n=tranwrd(&n,' D R ',");
&n=tranwrd(&n,' OD ',");
&n=tranwrd(&n,' O D ',");
&n=tranwrd(&n,' PC ',");
&n=tranwrd(&n,' P C ',");
&n=tranwrd(&n,' PA ',");
&n=tranwrd(&n,' P A ',");
&n=tranwrd(&n,' S C ',");
&n=tranwrd(&n,' CPA ',");
&n=tranwrd(&n,' CPAS ',");
&n=tranwrd(&n,' C P A ',");
&n=tranwrd(&n,' PARTNERS ',");
&n=tranwrd(&n,' PARTNER ',");
&n=tranwrd(&n,' SYSTS ',");
&n=tranwrd(&n,' SVC ',");
&n=tranwrd(&n,' SVCS ',");
&n=tranwrd(&n,' SLTNS ',");
&n=tranwrd(&n,' HOLDINGS ',");
&n=tranwrd(&n,' HOLDING ',");
&n=tranwrd(&n,' INTRNTL ',");
&n=tranwrd(&n,' U S ',' US ');
&n=tranwrd(&n,' PC ',");
&n=tranwrd(&n,' PTR ',");
&n=tranwrd(&n,' GEN ',");
&n=tranwrd(&n,' MBR ',");
&n=tranwrd(&n,' MGR ',");
&n=tranwrd(&n,' MEMBER ',");
&n=tranwrd(&n,' ET AL ',");
&n=tranwrd(&n,' SOLE ',");
&n=tranwrd(&n,' SINGLE ',");
&n=tranwrd(&n,' SNGL ',");
&n=tranwrd(&n,' OWNER ',");
a=index(&n,"INC ");if a=1 then &n=";drop a;
   if scan(&n,1)='U' and scan(&n,2)='S' then &n=tranwrd(&n,'U S ',' US ');
   if strip(&n)='PL' then &n=";
if strip(&n)='P L' then &n=";
```

```
if strip(&n)='PA' then &n='';
if strip(&n)='P A' then &n='';
if strip(&n)='HOLDINGS' then &n='';
if strip(&n)='ASSN' then &n='';
if strip(&n)='ASSOCS' then &n='';
if strip(&n)='LAW' then &n='';
if strip(&n)='SVCS' then &n='';
if strip(&n)='A PROFESSIONAL' then &n='';
if strip(&n)='FNDTN' then &n='';
if strip(&n)='L P' then &n='';
if strip(&n)='LP' then &n='';
if strip(&n)='CORP' then &n='';
if strip(&n)='INST' then &n='';
if strip(&n)='MGMT' then &n='';
if strip(&n)='LIABILITY' then &n='';
if strip(&n)='CO' then &n='';
if strip(&n)='INCOPORATED' then &n='';
if strip(&n)='INCORPARATED' then &n='';
if strip(&n)='INCORPERATED' then &n='';
if strip(&n)='INCORPORACTED' then &n='';
if strip(&n)='INCORPORADO' then &n='';
if strip(&n)='INCORPORATION' then &n='';
if strip(&n)='INCORPORAID' then &n='';
if strip(&n)='INCORPORATD' then &n='';
if strip(&n)='INCORPORATE' then &n='';
if strip(&n)='INCORPORATOR' then &n='';
if strip(&n)='INCORPORRATED' then &n='';
if strip(&n)='INCORPORTED' then &n=''; *strip all leading and trail blanks removed;
&n=strip(compbl(&n)); *reduce the spaces between words to one space;
   &s=tranwrd(&s,".","");
   &s=tranwrd(&s,",","");
   &s=tranwrd(&s,"& ","");
   &s=tranwrd(&s,"&","");
   &s=tranwrd(&s,"'","");
   &s=tranwrd(&s,"#","");
   &s=tranwrd(&s,"-","");
   &s=tranwrd(&s,"/","");
   &s=tranwrd(&s,"  "," ");
   &s=tranwrd(&s,"P O ","PO ");
   &s=tranwrd(&s," ZERO "," 0 ");
&s=tranwrd(&s," ONE "," 1 ");
&s=tranwrd(&s," TWO "," 2 ");
&s=tranwrd(&s," THREE "," 3 ");
```

```
&s=tranwrd(&s," FOUR "," 4 ");
&s=tranwrd(&s," FIVE "," 5 ");
&s=tranwrd(&s," SIX "," 6 ");
&s=tranwrd(&s," SEVEN "," 7 ");
&s=tranwrd(&s," EIGHT "," 8 ");
&s=tranwrd(&s," NINE "," 9 ");
&s=tranwrd(&s," TEN "," 10 ");
&s=tranwrd(&s," ELEVEN "," 11 ");
&s=tranwrd(&s," TWELVE "," 12 ");
&s=tranwrd(&s," THIRTEEN "," 13 ");
&s=tranwrd(&s," FOURTEEN "," 14 ");
&s=tranwrd(&s," FIFTEEN "," 15 ");
&s=tranwrd(&s," SIXTEEN "," 16 ");
&s=tranwrd(&s," SEVENTEEN "," 17 ");
&s=tranwrd(&s," EIGHTEEN "," 18 ");
&s=tranwrd(&s," NINETEEN "," 19 ");
&s=tranwrd(&s,"ZERO "," 0 ");
&s=tranwrd(&s,"ONE "," 1 ");
&s=tranwrd(&s,"TWO "," 2 ");
&s=tranwrd(&s,"THREE "," 3 ");
&s=tranwrd(&s,"FOUR "," 4 ");
&s=tranwrd(&s,"FIVE "," 5 ");
&s=tranwrd(&s,"SIX "," 6 ");
&s=tranwrd(&s,"SEVEN "," 7 ");
&s=tranwrd(&s,"EIGHT "," 8 ");
&s=tranwrd(&s,"NINE "," 9 ");
&s=tranwrd(&s,"TEN "," 10 ");
&s=tranwrd(&s,"ELEVEN "," 11 ");
&s=tranwrd(&s,"TWELVE "," 12 ");
&s=tranwrd(&s,"THIRTEEN "," 13 ");
&s=tranwrd(&s,"FOURTEEN "," 14 ");
&s=tranwrd(&s,"FIFTEEN "," 15 ");
&s=tranwrd(&s,"SIXTEEN "," 16 ");
&s=tranwrd(&s,"SEVENTEEN "," 17 ");
&s=tranwrd(&s,"EIGHTEEN "," 18 ");
&s=tranwrd(&s,"NINETEEN "," 19 ");
&s=tranwrd(&s," FIRST "," 1ST ");
&s=tranwrd(&s," SECOND "," 2ND ");
&s=tranwrd(&s," THIRD "," 3RD ");
&s=tranwrd(&s," FOURTH "," 4TH ");
&s=tranwrd(&s," FIFTH "," 5TH ");
&s=tranwrd(&s," SIXTH "," 6TH ");
&s=tranwrd(&s," SEVENTH "," 7TH ");
```

```
&s=tranwrd(&s," EIGHTH "," 8TH ");
&s=tranwrd(&s," NINTH "," 9TH ");
&s=tranwrd(&s," TENTH "," 10TH ");
&s=tranwrd(&s," ELEVENTH "," 11TH ");
&s=tranwrd(&s," TWELFTH "," 12TH ");
&s=tranwrd(&s," THIRTEENTH "," 13TH ");
&s=tranwrd(&s," FOURTEENTH "," 14TH ");
&s=tranwrd(&s," FIFTEENTH "," 15TH ");
&s=tranwrd(&s," SIXTEENTH "," 16TH ");
&s=tranwrd(&s," SEVENTEENTH "," 17TH ");
&s=tranwrd(&s," EIGHTEENTH "," 18TH ");
&s=tranwrd(&s," NINETEENTH "," 19TH ");
&s=tranwrd(&s," TWENTY "," 20 ");
&s=tranwrd(&s," THIRTY "," 30 ");
&s=tranwrd(&s," FORTY "," 40 ");
&s=tranwrd(&s," FIFTY "," 50 ");
&s=tranwrd(&s," SIXTY "," 60 ");
&s=tranwrd(&s," SEVENTY "," 70 ");
&s=tranwrd(&s," EIGHTY "," 80 ");
&s=tranwrd(&s," NINETY "," 90 ");
&s=tranwrd(&s,"TWENTY "," 20 ");
&s=tranwrd(&s,"THIRTY "," 30 ");
&s=tranwrd(&s,"FORTY "," 40 ");
&s=tranwrd(&s,"FIFTY "," 50 ");
&s=tranwrd(&s,"SIXTY "," 60 ");
&s=tranwrd(&s,"SEVENTY "," 70 ");
&s=tranwrd(&s,"EIGHTY "," 80 ");
&s=tranwrd(&s,"NINETY "," 90 ");
&s=tranwrd(&s," TWENTIETH "," 20TH ");
&s=tranwrd(&s," THIRTIETH "," 30TH ");
&s=tranwrd(&s," FORTIETH "," 40TH ");
&s=tranwrd(&s," FIFTIETH "," 50TH ");
&s=tranwrd(&s," SIXTIETH "," 60TH ");
&s=tranwrd(&s," SEVENTIETH "," 70TH ");
&s=tranwrd(&s," EIGHTIETH "," 80TH ");
&s=tranwrd(&s," NINETIETH "," 90TH ");
&s=tranwrd(&s," EXT ","");
&s=tranwrd(&s," STREET ","");
&s=tranwrd(&s," ST ","");
&s=tranwrd(&s," STE ","");
&s=tranwrd(&s," RD ","");
&s=tranwrd(&s," ST ","");
&s=tranwrd(&s,"BLVD","");
```

```
&s=tranwrd(&s," LN ","");
&s=tranwrd(&s," PL ","");
&s=tranwrd(&s,"PKWY","");
&s=tranwrd(&s,"PRKWY","");
&s=tranwrd(&s,"PWY","");
&s=tranwrd(&s,"FREEWAY","");
&s=tranwrd(&s,"FRWY","");
&s=tranwrd(&s,"FWY","");
&s=tranwrd(&s,"HIGHWAY","");
&s=tranwrd(&s,"HWY","");
&s=tranwrd(&s," STATE ","");
&s=tranwrd(&s," ROAD ","");
&s=tranwrd(&s," RR ","");
&s=tranwrd(&s," COUNTY ","");
&s=tranwrd(&s," CNTY ","");
&s=tranwrd(&s," US ","");
&s=tranwrd(&s," EXPRESSWAY ","");
&s=tranwrd(&s," EXPWY ","");
&s=tranwrd(&s," EXPY ","");
&s=tranwrd(&s," DR ","");
&s=tranwrd(&s," PIKE ","");
&s=tranwrd(&s," TERRACE ","");
&s=tranwrd(&s," TERR ","");
&s=tranwrd(&s," TER ","");
&s=tranwrd(&s," CT ","");
&s=tranwrd(&s," CIR "," ");
&s=tranwrd(&s," CTR ","");
&s=tranwrd(&s," WAY ","");
&s=tranwrd(&s," AVE ","");
&s=tranwrd(&s," PLZ ","");
&s=tranwrd(&s," PLAZA ","");
&s=tranwrd(&s," CROSSING ","");
&s=tranwrd(&s," XING ","");
&s=tranwrd(&s," FLOOR ","");
&s=tranwrd(&s," FLOO ","");
&s=tranwrd(&s," FLR ","");
&s=tranwrd(&s," FL ","");
&s=tranwrd(&s," LOT ","");
&s=tranwrd(&s," APT ","");
&s=tranwrd(&s," SUITE ","");
&s=tranwrd(&s," ROOM ","");
&s=tranwrd(&s," RM ","");
&s=tranwrd(&s," UNIT ","");
```

```
    &s=tranwrd(&s," UNITS ","");
    &s=tranwrd(&s," NUM ","");
    &s=tranwrd(&s,"#","");
    &s=tranwrd(&s," NO ","");
    &s=tranwrd(&s," BLDG ","");
    &s=tranwrd(&s," BLD ","");
    &s=compbl(&s);
    &s=strip(&s); *all leading and trailing blanks remove;
run;
%mend standard;


/*/ Do the loop /*/

%standard(Company, Address,Nets_&state.2007nodups)
```

Code Example 2: Code from Census in SAS to demonstrate matching with Business Register (This code requires the SAS Data Quality Server)

```
/* PROGRAM THAT MERGES the BUSINESS REGISTER (BR) for a particular year WITH NETS DATA */


/* The Business Register data set is establishment-based and includes business location, organization type (e.g., subsidiary or parent), industry classification, and operating data (e.g., receipts and employment).*/

/* The nets data has financial information (credit score, financial stress score) at the firm level. The only issue is that the majority of the observations in the nets data don't have an identifier that directly links the nets data with the business register database. In order to merge these two data sets we will do name and address matching */

/* THIS PROGRAM USES THE CLEAN BR (ONLY FOR CA AND DC) FROM THE
PROGRAM BR_CLEANUP AND CLEAN NETS DATA FROM THE PROGRAM NETS_CLEANUP */

options compress=yes;
libname NETS "; *path for the input files ;
run;
libname NETS2 "; * path for the output files - the merge data;
run;
LIBNAME br "; *the BR clean files;
run;

/*////////////////////////////////////////////////////////////////
/// Matching NETS to the BR by EIN (EMPLOYER IDENTIFICATION NUMBER)
////////////////////////////////////////////////////////////////*/
```

```
*defining macros;
%let yr=2005; *BR year to merge ;
%let year=05; *BR year to merge TWO DIGITS;
%let base=2005; *NETS year to merge;
%let st=CA; *state to merge;
%let statenum=06;

data netsfile;
	set nets.nets_&st._&base.; *nets file used in the merge;
	run;

data brfile;
 set br.br&yr.(where=(sstate="&st")); *br file used in the merge;
run;

data &mergefile &miss1(keep=nname nstreet ncity nzip nstate ) &miss2(keep=sname sstreet scity szip sstate);
	format nname nstreet ncity nzip sname sstreet scity szip;
	merge &out1(in=a) &out2(in=b);
	by ein;
	if a and b then output &mergefile;
	if a and not b then output &miss1;
	if b and not a then output &miss2;
run;


/* ////////////////////////////////////////////////////////
// NAME AND ADRESS MATCHING PART I
//////////////////////////////////////////////////////// */

/*
Step 1: We first match the NETS in year 1 with the BR in year 1, then we save the obs that we have not match from this first step and called this file unmatched1.
Step 2: We merge unmatched1 with the BR in year 0, then we save the obs we have not match from this second step and called this file unmatched2.
Step 3: We merge unmatched2 with the BR in year 2.

We proceed with these 3 steps to maximize the number of matches.
*/

%macro merge_current_post_pre(pre,current,post,statefips,statetwodigits);

%do iter=1 %to 3;
```

```
*re-defining macros;
%let base=&current.; *NETS year to merge;
%let st=&statetwodigits.; *state to merge;
%let statenum=&statefips.;

/* This files have been obtained from Nets_cleanup.sas and BR_cleanup.sas; */

* First iteration,merging BR year 1 with NETS year 1
%if &iter.=1 %then %do;
        %let yr=&base.; *BR year to merge ;
        %let year=%substr(&base.,3,2); *BR year to merge TWO DIGITS;
        data netsfile;
        set nets.nets_&st._&base.; *nets file use in the merge;
        run;
%end;

*Second iteration, merging (unmatched)NETS year 1 with BR year 0
%else %if &iter.=2 %then %do;
        %let yr=&pre.; *BR year to merge ;
        %let year=%substr(&pre.,3,2); *BR year to merge TWO DIGITS;
        data netsfile;
        set nets2.miss_nets10_nets&base._br&base._&st; * missing nets obs used in the previos merge;
        run;
%end;

* Third iteration, merging (unmatched)NETS year 1 with BR year 2
%else %if &iter.=3 %then %do;
        %let yr=&post.; *BR year to merge ;
        %let year=%substr(&post.,3,2); *BR year to merge TWO DIGITS;
        data netsfile;
        set nets2.miss_nets10_nets&base._br&pre._&st; * missing nets obs used in the previous merge;
        run;
%end;

data brfile;
 set br.br&yr.(where=(sstate="&st")); *br file use in the merge;
run;



/* ////////////////////////////////////////////////////////
// NAME AND ADRESS MATCHING PART II - DQmatching
//////////////////////////////////////////////////////// */
```

```
*Match everything using Proc DQ;
%DQLOAD (DQLOCALE=(ENUSA), DQSETUPLOC='sas pathname');

%macro match_name_address(match_number,s1,s2,input1,out1,matchcode,input2,out2,mergefile,miss1,miss2);
*two most important files here are input1 and input2 (the two tables that you want to merge);
*the rest are the outputs;
**********First match - Name and Address  **********;
proc dqmatch data=&input1 out=&out1 matchcode=&matchcode;
criteria
        var=nname matchdef='organization' sensitivity=&s1;
criteria
        var=nstreet matchdef='address' sensitivity=&s2;
run;

proc dqmatch data=&input2 out=&out2 matchcode=&matchcode;
criteria
        var=sname matchdef='organization' sensitivity=&s1;
criteria
        var=sstreet matchdef='address' sensitivity=&s2;
run;

proc sort data=&out1;
by &matchcode;
run;
proc sort data=&out2;
by &matchcode;
run;

data &mergefile &miss1(keep=nname nstreet ncity nzip nstate) &miss2(keep=sname sstreet scity szip sstate);
        format nname nstreet ncity nzip sname sstreet scity szip;
        merge &out1(in=a) &out2(in=b);
        by &matchcode;
        match=&match_number;
        if a and b then output &mergefile;
        if a and not b then output &miss1;
        if b and not a then output &miss2;
run;
%mend match_name_address;


options nomprint; *mprint;
*Pass1 ;
```

```
%match_name_address(1,80,80,netsfile,nets2.nets_match1,match_cd1,brfile,nets2.br_match1,nets2.mergepass1
,miss_nets,miss_br)
*Pass2 - The inputs for the second pass will be the missing matches from the first pass;
%match_name_address(2,80,55,miss_nets,nets2.nets_match2,match_cd2,miss_br,nets2.br_match2,nets2.merge
pass2,miss_nets2,miss_br2)
*Pass3 ;
%match_name_address(3,78,78,miss_nets2,nets2.nets_match3,match_cd3,miss_br2,nets2.br_match3,nets2.mer
gepass3,miss_nets3,miss_br3)

/* we could also chose to do more passes with different sensitivity, but in our case they were not good
matches*/

* Deleting data sets;
proc datasets library=nets2;
 delete Br_match: Nets_match: ;
quit;

**********Second match - Name and Zip**********;
%macro match_name_zip(match_number,s1,s2,input1,out1,matchcode,input2,out2,mergefile,miss1,miss2);
* Using what we did not match from nets or br we proceed;
proc dqmatch data=&input1 out=&out1 matchcode=&matchcode;
criteria
        var=nname matchdef='organization' sensitivity=&s1;
criteria
        var=nzip matchdef='text' sensitivity=&s2;
run;

proc dqmatch data=&input2 out=&out2 matchcode=&matchcode;
criteria
        var=sname matchdef='organization' sensitivity=&s1;
criteria
        var=szip matchdef='text' sensitivity=&s2;
run;
proc sort data=&out1;
by &matchcode;
run;
proc sort data=&out2;
by &matchcode;
run;

data &mergefile &miss1(keep=nname nstreet ncity nzip nstate) &miss2(keep=sname sstreet scity szip sstate);
        format nname nstreet ncity nzip sname sstreet scity szip; *ordering the variables in the data set;
        merge &out1(in=a) &out2(in=b);
```

```
        by &matchcode;
        match=&match_number;
        if a and b then output &mergefile;
        if a and not b then output &miss1;
        if b and not a then output &miss2;
run;
%mend match_name_zip;

*Pass4  ;
%match_name_zip(4,93,95,miss_nets3,nets2.nets_match4,match_cd4,miss_br3,nets2.br_match4,nets2.mergepa
ss4,nets2.miss_nets10,nets2.miss_br10)


data nets2.miss_nets10_nets&base._br&yr._&st;
set nets2.miss_nets10;
run;

 data nets2.miss_br10_nets&base._br&yr._&st;
set nets2.miss_br10;
run;

proc datasets library=nets2;
 delete Br_match: Nets_match: miss_nets10 miss_br10 miss_nets8 miss_br8;
quit;



**********Organize all of the Match files into One
File**********************************************;
proc datasets library=work;
 delete nets_br_match0;
quit;

proc append base=nets_br_match0 data=nets2.mergepass1 force;
run;
proc append base=nets_br_match0 data=nets2.mergepass2 force;
run;
proc append base=nets_br_match0 data=nets2.mergepass3 force;
run;
proc append base=nets_br_match0 data=nets2.mergepass4 force;
run;

/*
```

```
proc append base=nets_br_match0 data=nets2.mergepass5 force;
run;
proc append base=nets_br_match0 data=nets2.mergepass6 force;
run;
proc append base=nets_br_match0 data=nets2.mergepass7 force;
run;
proc append base=nets_br_match0 data=nets2.mergepass8 force;
run;
proc append base=nets_br_match0 data=nets2.mergepass9 force;
run;
proc append base=nets_br_match0 data=nets2.mergepass10 force;
run;
*/

proc datasets library=nets2;
 delete mergepass:;
quit;




/* This is the end of the matching, then we can merge our merge data sets to other data sets as the lbd */



%merge_current_post_pre(2002,2003,2004,06,CA);
```

Code Example 3: Code from CEO/DOL to implement probabilistic matching, EM algorithm (This code does NOT require the SAS Data Quality Server)

```
/* Select states and their corresponding ETA Regions*/
proc sql;
select state, reg into :st1-:st&sysmaxlong, :rg1-:rg&sysmaxlong
from ((select distinct(state), '6' as reg from pudf.WP_PY2013Q4_PUBSEEKERS_REG6)
union
(select distinct(state), '5' as reg from pudf.WP_PY2013Q4_PUBSEEKERS_REG5)
union
(select distinct(state), '4' as reg from pudf.WP_PY2013Q4_PUBSEEKERS_REG4)
union
(select distinct(state), '3' as reg from pudf.WP_PY2013Q4_PUBSEEKERS_REG3)
union
(select distinct(state), '2' as reg from pudf.WP_PY2013Q4_PUBSEEKERS_REG2)
union
```

```
(select distinct(state), '1' as reg from pudf.WP_PY2013Q4_PUBSEEKERS_REG1));
quit;

%let n=&sqlobs;

/* Selecting variables to be excluded for data linkage*/
proc sql;
select name into :var1-:var&sysmaxlong
from dictionary.columns
where memname="WP_PY2013Q4_PUBSEEKERS_REG1"
and upcase(name) not in ('ID','FILETYPE','STATE','PROGRAMYEAR','OBS','QUARTER',
        'BIRTH_DT', 'REG_DT','WIBNAME','GENDER','WIB','WHITE','INDIAN','ASIAN','BLACK',
        'MULTI','HAWAIIAN','HISPANIC','VET','VETELIG','VET911','VETCAMP','VETDIS',
        'VETTAP','VETRECENT','VETHOMELESS','VETTRANS','I','FIRST_SERVICE')
and libname='PUDF';
quit;

%let varnum=&sqlobs;

%macro EM;

/*Intializing p, u- and m-probabilities for the first iteration of the EM algorithm*/

/*Start state loop*/
%do s=1 %to &n;
        data kfs.em0&&ST&s;
        iter=0;
        p=.005;

  %do i=1 %to &varnum;
        m&&var&i=.9 ; u&&var&i=.1;
  %end;

        run;

        /*Start iteration loop*/
        %let j=0; %let stop=1;
        %do %until (&j=50 or &stop=0);
          %let j=%eval(&j+1);

        /*EM Algorithm*/
        data kfs.em&&ST&s;
```

```
	sysecho "Vector creation Iteration &j &&st&s &s of &n";

	set pudf.WP_PY2013Q4_PUBSEEKERS_REG&&RG&s (obs=1);

	declare hash ob(data set:%unquote(%bquote(')pudf.WP_PY2013Q4_PUBSEEKERS_REG&&RG&s
(where=(state="&&ST&s"))%bquote(')) , multidata:"Y", hashExp:16);
	ob.definekey('reg_dt','gender');
	ob.definedata(all:'Y');
	ob.definedone();

	set kfs.em0&&ST&s (where=(iter=%eval(&j-1)));
		n=0; n1=0; gm=0;

	DO UNTIL (eof);
		set PUDF.WP_PY2013Q3_SEEKERS_REG&&RG&s (where=(state="&&ST&s")
rename=(id=aid %do i=1 %to &varnum; &&var&i=a&&var&i %end;)) end=eof;

		n1+1;

		call missing(id %do i=1 %to &varnum; ,&&var&i %end;);
		rc=ob.find(key:reg_dt, key:gender);

		if rc=0 then do;
			m=1;u=1;

			%do i=1 %to &varnum;
			if m&&var&i>0 and u&&var&i>0 and m&&var&i>u&&var&i then do;
				if missing(&&var&i)=1 or missing(a&&var&i)=1 or &&var&i^=a&&var&i
				then do;
					m=m*(1-m&&var&i); u=u*(1-u&&var&i);
				end;
				else if &&var&i=a&&var&i then do;
					m=m*m&&var&i;
					u=u*u&&var&i;
				end;
			end;
			%end;

			g=round(p*m/(p*m+(1-p)*u),.00001);
			n+1;
			gm + g;

			%do i=1 %to &varnum;
```

```
if &&var&i=a&&var&i and missing(&&var&i)=0 and missing(a&&var&i)=0
then do;
        g&&var&i+1;

        if m&&var&i>0 and u&&var&i>0 and m&&var&i>u&&var&i then do;
                gm&&var&i + g;
                gu&&var&i + (1-g);
        end;

end;
%end;

rc=ob.has_next(RESULT: idother);
do while (idother ne 0);
        rc=ob.find_next(key:reg_dt, key:gender);
        rc=ob.has_next(result: idother);

        m=1;u=1;

        %do i=1 %to &varnum;
        if m&&var&i>0 and u&&var&i>0 and m&&var&i>u&&var&i then do;

                if missing(&&var&i)=1 or missing(a&&var&i)=1 or
                &&var&i^=a&&var&i then do;
                        m=m*(1-m&&var&i); u=u*(1-u&&var&i);
                end;
                else if &&var&i=a&&var&i then do;
                        m=m*m&&var&i;
                        u=u*u&&var&i;
                end;

        end;
        %end;

        g=round(p*m/(p*m+(1-p)*u),.00001);
        n+1;
        gm + g;

        %do i=1 %to &varnum;
        if &&var&i=a&&var&i and missing(&&var&i)=0 and missing(a&&var&i)=0
        then do;
                g&&var&i+1;
```

```
                        if m&&var&i>0 and u&&var&i>0 and m&&var&i>u&&var&i then do;
                            gm&&var&i + g;
                            gu&&var&i + (1-g);
                        end;

                    end;
                    %end;
                end;
            end;

            /*Calculating new p,u- and m-probability estimates based on EM algorithm*/
            if eof=1 then do;

                state="&&st&s";
                iter=&j;
                n2=ob.num_items;
                p=min(gm,n1,n2)/n;

                %do i=1 %to &varnum;
                if m&&var&i>0 and u&&var&i>0 and m&&var&i>u&&var&i then do;

                    if round(gm&&var&i/gm,.00001)=0 then m&&var&i=.00001;
                    else if round(gm&&var&i/gm,.00001)=1 then m&&var&i=1-.00001;
                    else m&&var&i = round(gm&&var&i/gm,.00001);

                    if round(gu&&var&i/(n-gm),.00001)=1 then u&&var&i=1-.00001;
                    else if round(gu&&var&i/(n-gm),.00001)=0 then u&&var&i=.00001;
                    else u&&var&i = round(gu&&var&i/(n-gm),.00001);

                end;

                u&&var&i = min(u&&var&i,round(g&&var&i/n,.00001));

                %end;

                keep state iter p n n1 n2 gm %do i=1 %to &varnum; m&&var&i u&&var&i %end;;
                output kfs.em&&ST&s;
            end;
end;
stop;
run;

%if &syserr ne 0 %then %do;
```

```
            %put ERROR: EM for &&st&s was not created;
        %let stop=0;
            %goto exit;
        %end;

        /*Checking for convergence of p, u- and m-probabilities
        between current and prior iterations*/

        proc sql noprint;
            select case when round(abs(a.m&var1-b.m&var1),.00001)
            %do i=2 %to &varnum; +round(abs(a.m&&var&i-b.m&&var&i),.00001) %end; =0
            or b.gm>min(b.n1,b.n2) then 0 else 1 end into :stop
            from kfs.em0&&ST&s a, kfs.em&&ST&s b
            where a.iter=%eval(&j-1);
        quit;

        %put For iteration &j stop was &stop;
        data kfs.em0&&ST&s;
        set kfs.em0&&ST&s kfs.em&&ST&s ;
        run;

        %exit:
        %end;
%end;
%mend;
%EM;
```

Code Example 4: Stata code from EBSA/DOL to remediate data quality issues and match data, using regular expressions

```
drop if accountant_firm_name=="" | accountant_firm_name=="0" | accountant_firm_name=="A"
drop if accountant_firm_ein=="" | accountant_firm_ein=="123456789" | accountant_firm_ein=="111111111" |
accountant_firm_ein=="000000000" |  accountant_firm_ein=="999999999"
*Drop plans with no-names
drop if regexm(accountant_firm_name,"PLEASE") & regexm(accountant_firm_name,"ATTACHMENT")
drop if regexm(accountant_firm_name,"SEE ATTACHMENT") | regexm(accountant_firm_name,"SEE
ATTACHED")
drop if regexm(accountant_firm_name,"TBD")
drop if regexm(accountant_firm_name,"TO BE DETERMINED")
drop if regexm(accountant_firm_name,"IN PROCESS")
drop if regexm(accountant_firm_name,"ACCOUNTANTS NAME")
drop if regexm(accountant_firm_name,"ABCDEFGHI")
drop if regexm(accountant_firm_name,"PDFDOC")
```

```
/*Cut down by EIN first, then assign the most common name in the data to that EIN*/
/*initial ID is the audit firm name*/
bysort accountant_firm_ein accountant_firm_name: gen num_names=_N
gsort accountant_firm_ein -num_names ack_id
bro accountant_firm_name accountant_firm_ein num_names
bysort accountant_firm_ein: gen num_ein=_N
bysort accountant_firm_ein: keep if _n==1
count

/*generate a shorter variable name*/
gen preproc=accountant_firm_name /*Preprocessed name*/
gen proccessed_name=accountant_firm_name
/*Standardize the Accountant Name*/
quietly replace proccessed_name = trim(proccessed_name)
quietly replace proccessed_name = upper(proccessed_name)
/*Remove funky characters*/
quietly replace proccessed_name=subinstr(proccessed_name,"(","",.)
quietly replace proccessed_name=subinstr(proccessed_name,")","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"%","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"#","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"""","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"`","",.)
quietly replace proccessed_name=subinstr(proccessed_name,","," ",.)
quietly replace proccessed_name=subinstr(proccessed_name,"{","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"}","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"[","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"]","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"\"," ",.)
quietly replace proccessed_name=subinstr(proccessed_name,"/"," ",.)
quietly replace proccessed_name=subinstr(proccessed_name,"&"," AND ",.)
quietly replace proccessed_name=subinstr(proccessed_name,"*","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"!","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"?","",.)
quietly replace proccessed_name=subinstr(proccessed_name,".","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"-","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"_","",.)

/*Doing this to remove any double spaces we will encounter from the / \ & replacements*/
quietly replace proccessed_name = itrim(proccessed_name)

/*Eliminate the most common words*/
quietly replace proccessed_name=subinstr(proccessed_name," AND ","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"ASSOCIATES","",.)
```

```
quietly replace proccessed_name=subinstr(proccessed_name," LLC","",.)
quietly replace proccessed_name=subinstr(proccessed_name," CPAS","",.)
quietly replace proccessed_name=subinstr(proccessed_name," PLLC","",.)
quietly replace proccessed_name=subinstr(proccessed_name," INC","",.)
quietly replace proccessed_name=subinstr(proccessed_name," LTD","",.)
quietly replace proccessed_name=subinstr(proccessed_name," LLP","",.)
quietly replace proccessed_name=subinstr(proccessed_name," CPA","",.)
quietly replace proccessed_name=subinstr(proccessed_name," PC","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"COMPANY","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"ASSOC","",.)
quietly replace proccessed_name=subinstr(proccessed_name,"FIRM","",.)

/*Get rid of the spacings*/
gen test=subinstr(proccessed_name," ","",.)
/*to eliminate the common letter combos as the end of strings, only take them off if they are the last two letters of the string*/
gen str_length=length(test)
gen lasttwo=substr(test,str_length-1,2)
replace test=substr(test,1,str_length-2) if inlist(lasttwo, "PA", "CO", "PC")

/*Only test on the first 20 characters of the auditor names*/
gen substring=substr(test,1,20)

/*This is the grouping substring function. I have set the matching threshold to be two edits for the longest string (2/20=.1)*/
/*Also prevents anything from shorter than 9 (1/9=.1111) to be matched if there are any differences*/
strgroup substring, gen(strgroupid) threshold(.175) first force
sort strgroupid
bro accountant_firm_ein accountant_firm_name strgroup
bysort strgroupid: gen flag_auditor_id=_n
count if flag_auditor_id==1
rename strgroupid summit_auditor_id
keep summit_auditor_id proccessed_name substring accountant_firm_ein
sort accountant_firm_ein
```

Code Example 5: Code from OSHA/DOL to normalize/canonicalize/remediate data quality issues. This uses regular expressions, and makes use of information from some fields that code example 1 deletes.

```
/*Name Standardization*/
data kfs.osha_reg5_HI;
set KFS.OSHA_REG5_HI;
If _n_=1 then do;
namekey= prxparse('s/\b(STORES?)? #?\d+$|\-|\\|\/|\.|\,|\bINC\b|\bINCORP[A-Z]+?\b|^THE\b|
\bCORP([A-Z]+)?\b|\bLLC\b|\bCOMPANY\b|\bCO\b|\bD.?B.?A.?\b|\bLTD\b|\(.+\)/ /I');
```

```
end;
retain namekey;
name=strip(compbl(compress(upcase(estab_name),"'")));
name = prxchange(namekey,-1,strip(name));
name = compbl(prxchange('s/ ?& ?|\sAND\s/ & /I',-1,name));
name = prxchange('s/\bUNITED STATES\b|\bU S\b/US/I',-1,name);
name = prxchange('s/\b(UNITED STATES|US) POSTAL SERVICE.*\b/USPS/I',-1,name);
name = prxchange('s/\bUNITED PARCEL SERVICE\b|\bU S\b/UPS/I',-1,name);
name = prxchange('s/\bSERVICES?\b|\bSRVCS\b/SRVC/I',-1,name);
name = prxchange('s/\bCENTERS?\b/CTR/I',-1,name);
name = prxchange('s/\bDEPARTMENT\b/DEPT/I',-1,name);
name = prxchange('s/\bHTLH/HEALTH/I',-1,name);
name = prxchange('s/\bHEALTH CARE/HEALTHCARE/I',-1,name);
name = prxchange('s/\b(SUPER)? ?MARKETS?\b/MARKET/I',-1,name);
name = prxchange('s/^USDOL OSHA.*/USDOL OSHA/I',-1,strip(name));
name = prxchange('s/^WAL.?MART.*/WALMART/I',-1,strip(name));
name = prxchange('s/^TYSON (\w+)? ?(FARMS?|FRESH|FOODS?).*/TYSON FOOD/I',-1,strip(name));
name = prxchange('s/^LOWES\b(OF|HOME|MILLWORK).*/LOWES HOME IMPROVEMENT/I',-
1,strip(name));
name = prxchange('s/^HOME ?DEPOT.*/HOME DEPOT/I',-1,strip(name));
name = prxchange('s/^FED( |-)EX.*/FEDEX/I',-1,strip(name));
name = prxchange('s/\bMANUF[A-Z]+/MFG/I',-1,strip(name));
name = prxchange('s/\bPACK[A-Z]+/PKG/I',-1,strip(name));
name = prxchange('s/.*\bA ?T & T\b.*/AT&T/I',-1,strip(name));
name = compbl(strip(name));
run;



/*Street Standardization*/
/*The majority of the following comes from a paper from a SAS Users Group meeting*/
data kfs.OSHA_REG5_HI;
set KFS.OSHA_REG5_HI;
If _n_=1 then do;
streetkey= prxparse('s/\sST\b|\sSTREET |\sAVE\b|\sAV\b|\sAVENUE\b|\sDR\b|\sDRIVE\b|
\sLN\b|\sLANE\b|\sRD\b|\sROAD\b|\sPKWY\b|\sPARKWAY\b|\sBLVD\b|\sBOULEVARD\b|\sPL\b|
\sPLACE\b|\sPLAZA\b|\sCT\b|\sCRT\b|\sCOURT\b|\sCIR\b|\sCIRCLE\b|\.|\-|\,|\(.+\)/ /I'); end;
retain streetkey;
street=strip(compbl(compress(upcase(site_address),"'")));
street = TRANWRD(street,'NORTH ','N ');
street = TRANWRD(street,'EAST ','E ');
street = TRANWRD(street,'WEST ','W ');
street = TRANWRD(street,'SOUTH ','S ');
```

```
street = prxchange('s/\bNORTHWEST\b|\bN W\b/NW/I',-1,street);
street = prxchange('s/\bNORTHEAST\b|\bN E\b/NE/I',-1,street);
street = prxchange('s/\bSOUTHWEST\b|\bS W\b/SW/I',-1,street);
street = prxchange('s/\bSOUTHEAST\b|\bS E\b/SE/I',-1,street);

street = prxchange(streetkey,-1,strip(street));
street = compbl(prxchange('s/ ?& ?|\sAND\s/ & /I',-1,street));
street = PRXCHANGE('s/#|STE|SUITE|BUILDING|BLDG/ ZTE /',-1,street);
street = PRXCHANGE('s/FRWY|FREEWAY|FWY/ FWY /',-1,street);
street = PRXCHANGE('s/EXPRWY|EXPRESSWAY|EXPWY|EXPY/ EXPY /',-1,street);
street = PRXCHANGE('s/HIWAY|HIGHWAY/ HWY /',-1,street);
street = PRXCHANGE('s/P ?(0|O) ?BOX\b|\bPMB\b|\bP ?O DRAWER\b|
\bPOST OFFICE DRAWER\b|\bDRAWER\b/ ZPB /',-1,street);
street = PRXCHANGE('s/ RM | ROOM / ZRM /',-1,street);
street = TRANWRD(street,'FIRST ','1ST ');
street = TRANWRD(street,'SECOND ','2ND ');
street = TRANWRD(street,'THIRD ','3RD ');
street = TRANWRD(street,'FOURTH ','4TH ');
street = TRANWRD(street,'FIFTH ','5TH ');
street = TRANWRD(street,'SIXTH ','6TH ');
street = TRANWRD(street,'SEVENTH ','7TH ');
street = TRANWRD(street,'EIGHTH ','8TH ');
street = TRANWRD(street,'NINTH ','9TH ');
street = TRANWRD(street,'TENTH ','10TH ');
street = TRANWRD(street,'ONE ','1 ');
street = TRANWRD(street,'TWO ','2 ');
street = TRANWRD(street,'THREE ','3 ');
street = TRANWRD(street,'FOUR ','4 ');
street = TRANWRD(street,'FIVE ','5 ');
street = TRANWRD(street,'SIX ','6 ');
street = TRANWRD(street,'SEVEN ','7 ');
street = TRANWRD(street,'EIGHT ','8 ');
street = TRANWRD(street,'NINE ','9 ');
street = TRANWRD(street,'TEN ','10 ');
street = PRXCHANGE('s/TWENTY.?(?=[0-9])/2/I',-1,STREET);

run;

/*This macro is not mine but thought I'd share. It likely needs extensive customization and review for individual
needs*/
%MACRO BREAKUP_ADD (PATTERN=,VAR=,NUM=,NEWVAR=);
IF _N_=1 THEN DO;
RETAIN ExpID&NUM;
```

```
PATTERN="/&PATTERN/I";
ExpID&NUM=PRXPARSE(PATTERN);
END;
CALL PRXSUBSTR(ExpID&NUM, &VAR, POSITION&NUM);
/*The first half of the macro creates a DO Loop that looks for the patterns in the addresses that were created in the
first macro, e.g. ZRM. When the pattern is found, it outputs the starting position of the pattern.*/
IF POSITION&NUM = 1 THEN DO;
MATCH = SUBSTR(&VAR,POSITION&NUM);
&NEWVAR=MATCH;
END;
IF INDEX(&VAR,"&PATTERN") THEN &NEWVAR=SUBSTR(&VAR,POSITION&NUM);
IF INDEX(&VAR,"&PATTERN") THEN SUBSTR(&VAR,POSITION&NUM)=";
%MEND BREAKUP_ADD;
data kfs.OSHA_REG5_HI (DROP=POSITION1 POSITION2 POSITION3 POSITION4 POSITION5
ExpID1 ExpID2 ExpID3 ExpID4 ExpID5 MATCH PATTERN);
set KFS.OSHA_REG5_HI;
%BREAKUP_ADD (PATTERN =ZPB,VAR=STREET,NUM=1,NEWVAR=PO_BOX_STREET);
%BREAKUP_ADD (PATTERN=ZTE,VAR=STREET,NUM=2,NEWVAR=SUITE_STREET);
%BREAKUP_ADD (PATTERN=ZTE,VAR=PO_BOX_STREET,NUM=3,NEWVAR=SUITE_STREET);
%BREAKUP_ADD (PATTERN=ZRM,VAR=STREET1,NUM=4,NEWVAR=RM_STREET);
%BREAKUP_ADD (PATTERN=ZRM,VAR=SUITE_STREET,NUM=5,NEWVAR=RM_STREET);
PO_BOX_STREET = STRIP(TRANWRD(PO_BOX_STREET,'ZPB',''));
SUITE_STREET = STRIP(TRANWRD(SUITE_STREET,'ZTE',''));
RM_STREET = STRIP(TRANWRD(RM_STREET,'ZRM',''));
PO_BOX_STREET=COMPRESS(PO_BOX_STREET,'#');
SUITE_STREET=COMPRESS(SUITE_STREET,'#');
RUN;
```

Code Example 6: Code from CEO/DOL to composite string distance measures

Code file in R, along with a sample data file, for evaluating the relative effectiveness and contrasts between common string distance measures. The code also demonstrates composite string distance measurement, and using high levels of similarity and dissimilarity to identify data quality issues. The code and data file can be found at the following address:
https://github.com/dullandboring/employer-data-matching

Code Example 7: Code compiled by Rebecca Bilbro, in Python, to test string matching functions

The following page contains a variety of tools and code examples to demonstrate and test different string distance measures. The resources include native capabilities in the base distribution, as well as DIFFLIB, FuzzyWuzzy, Jaccard and Jellyfish modules. The code for these varying examples, libraries and functions can be found at the following address:
https://github.com/DruidSmith/Python-Matching-Algorithms/blob/master/String%20Comparison.ipynb

Code Example 8: Module used by OFLC, ETA/DOL to remediate data quality issues
The following code is used by the ETA Office of Foreign Labor Certification (OFLC) to standardize string content in order to minimize the amount of manual de-duplicating required when trying to match or aggregate data.

```
Option Compare Database

Sub ModifyData()

DoCmd.SetWarnings False

'Convert string to proper case: UPDATE PW SET PW.EMPLOYER_LEGAL_BUSINESS_NAME = StrConv([EMPLOYER_LEGAL_BUSINESS_NAME],3)

DoCmd.OpenQuery "UPDATE_Employer_Name_Proper_Case"


'Remove Special Characters from a string: UPDATE PW SET PW.EMPLOYER_LEGAL_BUSINESS_NAME = Replace([EMPLOYER_LEGAL_BUSINESS_NAME],"enter special character between quotes"," ")

DoCmd.OpenQuery "UPDATE_Remove_Ampersands"

DoCmd.OpenQuery "UPDATE_Remove_Periods"

DoCmd.OpenQuery "UPDATE_Remove_Commas"

DoCmd.OpenQuery "UPDATE_Remove_Semicolons"

DoCmd.OpenQuery "UPDATE_Remove_Colons"

DoCmd.OpenQuery "UPDATE_Remove_Dashes"

DoCmd.OpenQuery "UPDATE_Remove_Underscores"

DoCmd.OpenQuery "UPDATE_Remove_Open_Paren"

DoCmd.OpenQuery "UPDATE_Remove_Closed_Paren"

DoCmd.OpenQuery "UPDATE_Remove_Apotrophe"

DoCmd.OpenQuery "UPDATE_Remove_Additional_Spaces"

DoCmd.SetWarnings True

MsgBox "Complete"

End Sub
```