DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 137–165, http://dx.doi.org/10.2478/JOS-2019-0007

An Evolutionary Schema for Using "it-is-what-it-is" Data in Official Statistics

Jack Lothian[1], *Anders Holmberg*[2], *and Allyson Seyb*[3]

The linking of disparate data sets across time, space and sources is probably the foremost current issue facing Central Statistical Agencies (CSA). If one reviews the current literature looking for the prevalent challenges facing CSAs, three issues stand out: 1) using administrative data effectively; 2) big data and what it means for CSAs; and 3) integrating disparate data set (such as health, education and wealth) to provide measurable facts that can guide policy makers. CSAs are being challenged to explore the same kind of challenges faced by Google, Facebook, and Yahoo, which are using graphical/semantic web models for organizing, searching and analysing data. Additionally, time and space (geography) are becoming more important dimensions (domains) for CSAs as they start to explore new data sources and ways to integrate those to study relationships. Central agency methodologists are being pushed to include these new perspectives into their standard theories, practises and policies. Like most methodologists, the authors see surveys and the publications of their results as a process where estimation is the key tool to achieve the final goal of an accurate statistical output. Randomness and sampling exists to support this goal, and early on it was clear to us that the incoming "it-is-what-it-is" data sources were not randomly selected. These sources were obviously biased and thus would produce biased estimates. So, we set out to design a strategy to deal with this issue.

This article presents a schema for integrating and linking traditional and non-traditional datasets. Like all survey methodologies, this schema addresses the fundamental issues of representativeness, estimation and total survey error measurement.

Key words: Representativeness; timeline databases; statistical registers; Estimation; administrative data.

1. Introduction

The linking of disparate data sets across time, space and sources is probably the foremost current issue facing Central Statistical Agencies CSAs. If one reviews the current literature looking for the prevalent challenges facing CSAs, three issues stand out: 1) using administrative data effectively; 2) big data and what it means for CSAs; and 3) integrating disparate data sets (such as health, education and wealth) to provide measurable facts that can guide policymakers. CSAs are being challenged to explore the same kind of concerns facing Google and Facebook, which are using graphical/semantic web models Ferrara et al. 2011 for organizing, searching and analyzing data. Additionally, time and space

[1] 360 Hinton Ave S, Ottawa ON K1Y1A5 Canada. Email: lothianjack@netscape.net
[2] Statistics Norway, Division for Methodology, Akersveien 26 Oslo, Norway. Email: anders.holmberg@ssb.no
[3] Stats NZ, Statistical Methods, Private Bag 4741, Christchurch 8011, New Zealand. Email: allyson.seyb@stats.govt.nz

(geography) are becoming more important dimensions (domains) for CSAs as they start to explore causal models. Central agency methodologists are being pushed to include these new perspectives into their standard theories, practises and policies. This article presents a schema for integrating and linking traditional and nontraditional data sets. Like all survey methodologies, this schema addresses the fundamental issues of representativeness, estimation and total survey error measurement.

Over the past decade, a new design paradigm has emerged concerning strategies for integrating disparate data sets to provide new understandings from the data. The development of this paradigm is currently not focused and there are multiple paths of advancement being pursued, such as big data, evolutionary databases (Fowler and Sadalage 2003), semantic web models, graphical query databases and many others. In a Graphical Queries Database (GQD), every element has a direct pointer to its adjacent elements. The simplest GQD pointer is a first order tree of one-to-one links. Semantic web links are first-order trees where the linkage function (the verb) becomes a generalized function. All these areas of research have a core issue: combining/linking large disparate data sets in a feasible and cost-effective manner. The complexity of the information, the fuzziness of the data inside each data set, the fuzziness of the linkage strategies, the large number of disparate data set, covering disjoint populations, the lack of control of the content and quality of administrative data, and the size of the data sets preclude the use of many straightforward classical solutions (Baker et al. 2013; Bakker and Daas 2012; Hand 2018; Holt 2000; Zhang 2012).

All the above-mentioned strategies appear to follow parallel paths to the same general solution. All these approaches propose viewing disparate data set integration as an evolutionary or ongoing process. The data and the database structure evolve as new information is added; as more data sets are added and linked; as new relationships between data sets are discovered and added; as new models of how different data sets interact are discovered; as new editing rules and methodologies are found; as the questions that we want answered change; as we become more knowledgeable of the data; and so on. The evolutionary nature of the problem implies that no fixed solution can succeed over an extended period. All the strategies cited above embrace evolution and make it part of the solution.

The core data design concept we are proposing is a simplified adaption of how many online search companies structure and search data. The major point of departure of our schema versus these online solutions is our inclusion of time. For these companies, the point in time at which the measurements are made is not usually a relevant characteristic, but for our schema, it will be a fundamental aspect of the data. Later in the document, we will also see that “space” or geography will become a necessary dimension of our design schema. Our schema will be underpinned and anchored by a space-time lattice, through which our entities will travel. It is somewhat akin to the game called “Life”. We will call the structured collection of common files (administrative, survey, register or census) an evolutionary schema. The term “evolutionary” implies that the database constructs entities’ event timelines and these timelines are updated with new current events. The event timelines evolve. In the paradigm of database design and programming, “evolutionary database” design has a different sense. It is the database design schema and algorithms that are always evolving in an incremental fashion. Our proposed design will be evolutionary in this sense as well.

We present a conceptual schema for dealing with the integration of nontraditional and traditional survey data sets. It is important to note that we will be presenting a strategy for structuring, analyzing problems and answering questions, rather than a specific solution. As in classical survey design, our final goal will be a strategy to provide the best possible estimates. To achieve this, we convey the message that methodologists must understand the whole process that will produce the estimates, not just focus on one phase of the process.

We believe that the basis for understanding this process and creating interpretable and meaningful estimates will be a system of statistical base registers, plus consistent monitoring and maintenance strategies. These statistical registers serve as lighthouses for illuminating 'trusted' estimation procedures and provide a benchmark for comparing and investigating representativeness concerns. We believe that our schema provides a broad and general framework for CSAs working with large collections of administrative data and other conveniently available data sets/databases that we refer to as "it-is-what-it-is" data sets. We offer a framework for: structuring the non-probabilistic data; making it useful for cause and effect statistical inference; incrementally developing, designing and maintaining the database system; and, inserting total survey error concepts into the schema. Our schema does not provide detailed designs for these processes, instead we provide a pseudo-scientific framework for addressing survey design questions when using non-probabilistic data sets.

Section 2 presents the concept of "it-is-what-it-is" data sources. Section 3 discusses the importance of registers to support estimation and eliminate potential biases within the database schema. Section 4 presents an overview of our data model for structuring and using the data in the evolutionary database. Section 5 discusses how estimation might take place in the evolutionary schema. Section 6 discusses the place of metadata, and measuring Total Survey Error (TSE) and controlling quality in this evolutionary schema. Section 7 is a summary.

2. Structured Framework for Using "it-what-it-is" Data Sets

2.1. "It-is-what-it-is" Data Sets

In our evolutionary schema, we will assume that all data sources integrated into the evolutionary database are provided by an outside agency that is beyond the control and influence of the owners of the evolutionary database. These outside sources could be administrative files, censuses, registers, client lists, commercial transactions, sensor readings, survey files, sample files, and so on. Our source data sets will be what Sharon Lohr (Lohr et al. 2015) recently referred to as "it-is-what-it-is" data sets. "It-is-what-it-is" data sets are source files where the survey methodologist has no control over the selection probabilities, nor the content of the files. It should be noted that the true sample selection probabilities for the entities in these external sourced data sets may be non-probabilistic and/or unknowable.

As expressed by Sharon Lohr, the term "it-is-what-it-is" has a wider sense. As survey methodologists, we may be asked to answer questions where the sole source of information concerning these questions are "it- is-what-it-is" data sets and thus we may be forced to use these data sets despite their limitations. In this case, methodologists must resort to pseudo-scientific methods to address the questions. If this is the case, Sharon says methodologists need to be aware of the data sets' limitations or "what it is". Sharon stated that "it-is-what-it-is" data sets fundamentally change our analysis paradigm and we need to understand this point. In the following discussions, the "it-is-what-it-is" nature of the data sources will be an integral part of our schema.

In our case, both administrative data and big data fit the "it-is-what-it-is" concept. They are not necessarily distinct from one another, and from a CSA's perspective, using them means reusing data that originated outside the agency. UNECE (2011) defines administrative data as "data that is collected by sources external to statistical offices" and "administrative sources are data holdings containing information that is not primarily collected for statistical purposes". This broad definition would also include almost all big data, given the existing different definitions of the phenomena. However, an administrative data source does not have to be "big" nor do big data sources normally have administrative purposes. Usually, the administrative data delivered to the CSAs come from and through the operations of another public organization. This is seldom the case with big data sources; they stem from activities, events and operations within the whole of society.

2.2. *The "Elephant in the Room" – Representativeness*

Unfortunately, there is an elephant in the room when we deal with "it-is-what-it-is" data sets. The elephant is the fact that these source files may not be appropriate for making statistical inferences concerning the general population because the selection probabilities are non-probabilistic.˙ A recent American Association Public Opinion Research (AAPOR) task force report on non-probability sampling (Baker et al. 2013) stated that "approaches lacking a theoretical basis are not appropriate for making statistical inferences". It was pointed out in an earlier AAPOR report (Baker et al. 2010) that statistical estimates and inferences drawn from "it-is-what-it-is" data sets cannot be trusted to be representative of the general population. In reference to the two AAPOR reports, Langer (Langer 2013) quotes a well-known classical reference (Kruskal and Mosteller 1979) stating that "[w]e prefer to exclude non-probability sampling methods from the representative rubric." Langer is implying that one cannot ever claim that results derived from an "it-is-what-it-is" data set are representative of the general population. This is a strong statement and raises questions about the ultimate usefulness of "it-is-what-it-is" data sources.

As the 2013 AAPOR report states, the key is the risk associated with the source data set not being representative of the general population. This is a serious risk because most "it-is-what-it-is" sources suffer from significant coverage issues associated with various sub-populations within a target population. (Overcoverage, duplicates, undercoverage, and missing data can occur in any data source and they can all lead to a population or sample being nonrepresentative. For brevity, at times we will use these terms interchangeably or in a generic sense.) The risk of bias is a systemic problem when dealing with "it-is-what-it-is" data sets and, as illustrated in Subsection 4.4, the cross-linking of "it-is-what-it-is" data sets significantly increases the potential risk. This is the Achilles heel of "it-is-what-it-is" data and, if we cannot address this issue, we will never be able to widely use "it-is-what-it-is" data. As survey methodologists, we must be able to defend ourselves from criticisms of bias caused by, for example, undercovering populations, such as the underprivileged or rare populations. Without a methodology to measure coverage issues and correct its effects, how do we maintain our credibility? Our schema offers a strategy for confronting this key issue and a stepping stone enabling CSAs to handle a paradigm change and make statistics by repurposing and combining data from sources outside their direct control.

2.3. Correcting Nonrepresentativeness with Registers and Frames

In our article, we address the representativeness risk by creating statistical population registers or frames that allow us to measure and correct over- and undercoverage. Most CSAs estimate for three types of populations: persons, businesses (from a National Accounts perspective these include nonprofit and public organizations) and geography. So, we propose that CSAs adopt these registers as their fundamental mechanism for dealing with nonrepresentativeness within their data ecosystem. As we go through the next few sections we will outline a strategy that:

1. Creates three lighthouse (base) registers systems and uses them to measure under- and overcoverage in various strata. Then, we use these registers to create calibrations (design-based designs) or models (model-based designs) or Bayesian priors (Bayesian designs) to correct for under- and overcoverage.
2. We will assume that we can construct a stratification definition process that ensures that within each stratum we can assume that the observed entities were generated by a random process. Thus, within each stratum the observed entities are assumed to be representative of the stratum sub- population.
3. This estimation capability will be supported by efficient and frequent monitoring of entity transitions in the registers. We foresee the regular use of indicators of entity flows and means of validation (through surveys and investigations) that are regularly used to update the strata and tombstone information in the registers.

The authors recognize that their strategy is naive and pseudo-scientific. It may not correct for all the biases created by the "it-is-what-it-is" data sets' under- and overcoverage. Yet it is a first step along a well-trodden design-based path. As we gather more expertise, future methodologists will develop more mature and complex methodologies for dealing with representativeness. By anchoring "it-is-what-it-is" data over time against better known or controlled data, there will be progress. While we recognize the risks of following a strategy that does not have an unambiguous theory behind it, we feel there is no other choice.

3. Creating the Lighthouse Registers

3.1. Estimation and Representativeness Requirements Imply the Need for Registers

Our evolutionary schema's estimation strategy is built upon three lighthouse registers systems (Thygesen and Grosen-Mielsen 2013). The structural supports for estimation will be the three traditional entity registers/frames already used by many CSAs. These are geography (or land, dwellings, property, or addresses), persons (or households, or families) and firms (or organizations, businesses, enterprises, establishments, or plants). Traditionally, the census played the role of both the dwelling and person registers, while the business register provided a firm register. To these three key base registers we add time, so that cause and effect relationships can be studied. As in science, time is a special dimension with unique properties quite different from our other three entity dimensions. Yet nevertheless, it will enter many estimation problems. Note that in our schema, a base register system is not equivalent to a sampling frame nor to a census. A census could be an input for building a register and a frame is an output of a register system. In practice, base register systems might be a single database file spanning all time periods and subpopulations or it might be a collection of subregistries achieving the same purpose. Base registers: define important statistical units, define standardized populations, contain links to units in other base registers, contain links to other data sources that relate to the same units, are important as a sampling frame, and can be used for demographic statistics for the units (Wallgren and Wallgren 2014). For convenience, we will dispense with using the descriptor "system" and use the singular form "register" when referring to base register systems.

The base registers of geography (LR), persons (PR) and businesses (BR) together with time are the lighthouses for estimation in their respective dimension. They illuminate potential areas of bias in our estimation system and shine light on the quality of our estimates (Figure 1). Base registers that are connectable to our data sources are the key design element that will allow us to make high quality estimations. These registers will be our starting point for adjusting for nonrepresentativeness effects in our schema.

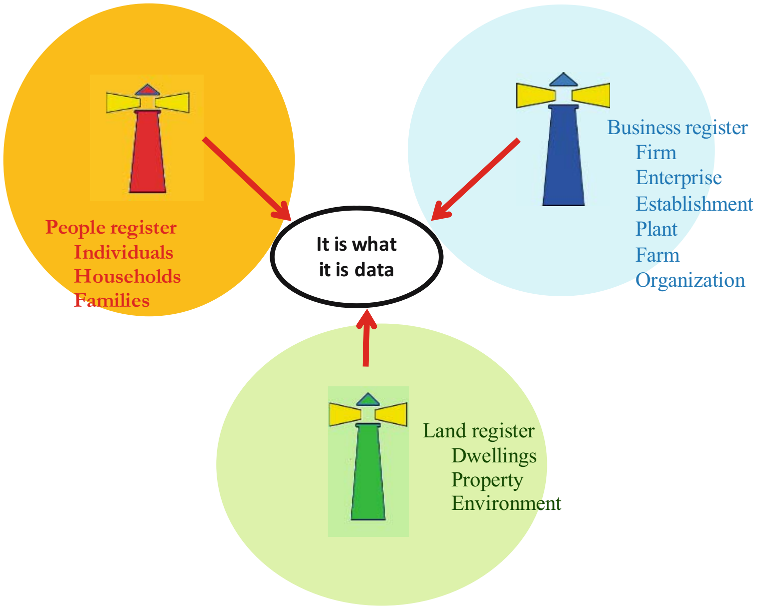

Fig. 1. The three base registers anchor estimates and illuminate the quality of "it-is-what-it-is" data sets.

Historically CSAs have built business registers from administrative files, while censuses and administive address files have been the main sources to make pseudoregisters for dwellings and persons. We envision a future where all three base registers are built from administrative files supplemented by surveys and censuses.

The challenge is how to create a base register from a disparate collection of administrative files. If we create a union data set of all the entities covered by multiple administrative data set sources, the resulting population size is often orders of magnitude larger than the current estimated population count of entities. Alternatively, when we create an intersection data set for the available administrative data sources, the coverage of the population rapidly plummets as more administrative data sets are joined. Typically, the entity count in the intersection data set is much lower than the current estimated population of entities. Winnowing down the number of entities in this collective to create an unbiased frame with complete unduplicated coverage of a desired target is a complex task. We believe that this implies that we must recognize representativeness as the core issue in developing our register systems. The data sources will be "it-is-what-it-is" data sets and we accept that fact and deal with it the best we can. Without some structured strategy for dealing with this issue, any estimation process will be of poor quality. Fortunately, CSAs have a template for developing these future registers: the current universally accepted process for creating business registers.

In the following sections, we will outline a strategy for constructing a representative base register. There are commonalities of function and design that cross the three base registers defined within our schema. Each register will be made up of well-defined entities that exist in the real world and are theoretically finite in number and countable. In each case, the registers will cover a wider population of entities than the current active population. The register will be created from multiple sources of varying quality, indicating whether an entity existed or not and over which period. Entities will have birth and death dates. Hierarchical structures may exist within each register: for businesses (enterprises, establishments, plants, locations); for persons (households, families, persons); and geography (regions, census tracts, land holdings, buildings, addresses).

3.2. Universal Identifiers

Most of the literature on creating person registers was written by authors from countries where universal and unique personal identifiers exist and have been in use by the general population for decades (Bakker and Daas 2012). If a country has a universal identifier given to every resident at birth or upon entering the country for residence, then they have the core of a personal registry. Yet even in this fortunate case, this will not be sufficient information for creating a base person register. To create a PR, one also needs entry and exit dates from the country and birth and death dates for every entity. Alternatively, a country that conducts regular censuses has the core for creating a person register. But in this case, they must coherently merge all the available censuses into a unified file, plus augment the information from other sources detailing births and deaths, and entries and exits to the country. Adding time to our schema complicates our design.

Outside of Europe, universal identifier registers are rare. In most countries, there are serious technical and political obstacles to overcome if one wishes to create a universal identifier system. The authors believe that for the foreseeable future, most of the countries in the world will not have a universal person identifier system. Therefore, a generic register construction methodology should not depend on the existence of a universal identifier. This requirement presents us with a conundrum if neither a universal identifier system nor regular censuses exists. We are confident that the current available BR technology, together with an evolutionary development strategy can overcome this challenge. In the following sections, we will assume that a universal identifier does not exist.

3.3. A Template for a Base Register

The basic BR template is mature, well understood and supported by a broad international consensus. Thus, we will use a simplified BR structure as a template for illustrating how one might construct and maintain the three base registers. In common usage, the term "business register" is not a base register system in the sense that we define it. The current standard design of the business register encompasses a complex system of interacting files, rather than one core list. In our schema, we define three types of files encompassed within each base register system: the source files or *administrative registers;* the *entity register* which is our core base register; and the *statistical registers,* which we might think of as statistical frames.

The BR's *entity register* of firms enumerates all business entities that have an event in any "it-is-what-it-is" source data files in the past Ω years. These collections of source files used to identify and birth entities to the entity register we will call the *administrative registers* and are subregisters within the business register system. The BR's administrative registers can give conflicting or partial information concerning the presence of business activity at any point in time. They can have different processing dates with different lags in their arrival at a CSA. Different data source agencies tend to use different identifiers. Duplicate transactions can occur. The information collected by each source can be radically different, with conflicting evidence concerning events and activity. Thus, the entities birthed from the administrative registers into the *entity register* represent a spectrum of entities with a varying quality of information. Some entities will have definitive birth dates and continuous ongoing economic activity over a span of time. Others may only show evidence of an entity registration, with no sign of subsequent activity. We divide this spectrum into three groups. The groupings depend on the quality of the administrative information indicating whether the entity exists and is active in a specific target population at a specific time. At the top of the spectrum are entities with multiple confirmed indicators of existence from high quality sources and at the bottom of the spectrum are firms with only partial information from one low quality source.

All sources are not considered of equal quality or informative value concerning the presence of activity. As such, rules must be created that define when the activity observed implies an actual birth to the register. When maintaining the register, there is a trade-off concerning the breadth of information included in the entity register versus the cost of processing and maintenance of this information. Typically, a small number of sources are viewed as "fundamental" indicators. New registrations from these fundamental sources will always generate births to the base register. Internal IDs for these fundamental sources are maintained within the entity register. Typically, these maintenance processes ensure that the internal IDs are consistent over time and space, and are unduplicated. If there are multiple IDs on the registers, there should exist an internally generated unique ID that spans the population of all the fundamental sources. If hierarchical structures exist on the register, then multiple internal IDs may be required. The entity register can include "inactive" entities: entities that show uncertain, infrequent, very weak or no signal of activity. (For business entities, "inactive" is an acceptable terminology, but for the PR it would be inappropriate. A preferred terminology then might be a classification of "unconfirmed".) An example of an inactive firm is a single indicator of registration or creation on a single administrative file with no known other event. Figure 2 summarizes this process.

The full register system is not useable in any real sense by users outside CSAs because it contains a collage of disparate populations and plethora of active and inactive entities. The essential process in Phase 1 is the "birthing rules" shown in Figure 2. In the BR, these rules are typically set by a cooperative team from the BR, National Accounts, business surveys and methodology. External users of the register cannot change or control these rules.

The administrative and entity registers are never seen by external users. Instead, the users see the *statistical registers* (or target population frames). A frame is the empirically derived list of the target population of interest. In the design-based paradigm, it would be our sampling frame and most practitioners think of a frame in the context of a sampling but, it is also our best possible estimated entity count. The register system presents statistical registers (or survey frames) to users by putting a filter over the full register. Each filter changes what entities the user will see. This is the second phase of the register maintenance system. Figure 3 below summarizes this process. The key process in phase 2 is the "statistical register creation rules".

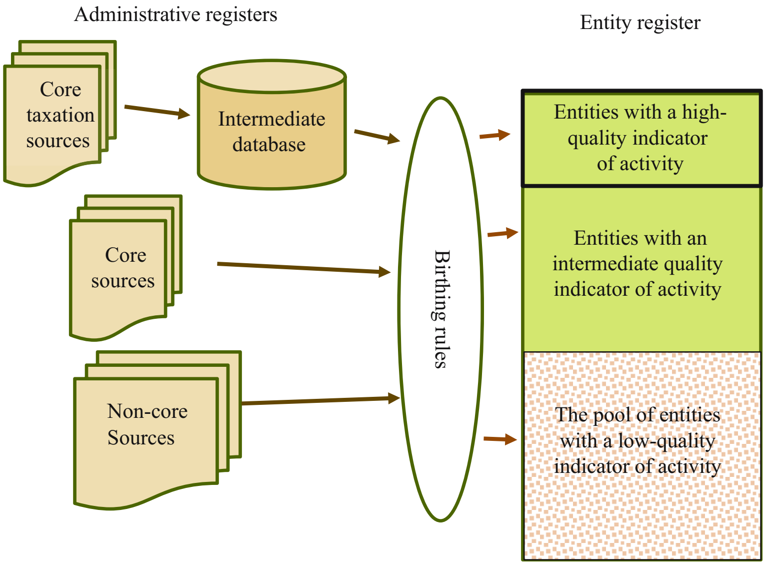
Administrative registers
Entity register
Core taxation sources
Intermediate database
Core sources
Non-core Sources
Birthing rules
Entities with a high-quality indicator of activity
Entities with an intermediate quality indicator of activity
The pool of entities with a low-quality indicator of activity

Fig. 2. Phase 1: The birthing and maintenance processes for generating an entity register.

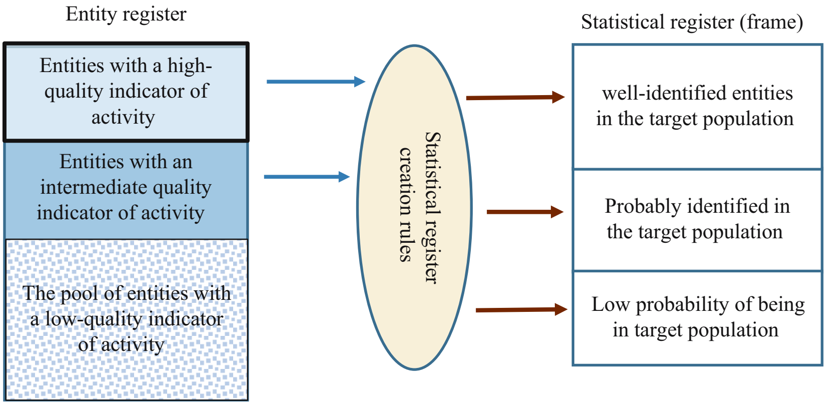
Entity register
Entities with a high-quality indicator of activity
Entities with an intermediate quality indicator of activity
The pool of entities with a low-quality indicator of activity
Statistical register creation rules
Statistical register (frame)
well-identified entities in the target population
Probably identified in the target population
Low probability of being in target population

Fig. 3. Phase 2: Creation of statistical register and target population frame.

The real output of the registers is frame lists from the BR, PR, and LR of target populations at a given time for example, lists of active firms, resident populations or property holdings. In our paradigm, it is these *statistical registers* that become our lighthouses for identifying and decreasing potential biases.

The "statistical register creation rules" are a filtering mask that uses standardized rules to extract statistical registers. Users will have no control over the definitions of the standardized filtering fields, but they will have considerable flexibility within these definitions. In the case of the BR, the user may extract any standard industrial code (SIC) grouping, but they will not be able to change SIC definitions. Similarly, BR users would not be permitted to change the definition of "inactive firms" but they could choose to select active and/or inactive firms. Of course, inactive firms will not be maintained to the same standards as active firms.

Typically, a CSA will generate a full business frame (statistical business register) every publication period (monthly, quarterly or annually). Thus, a design principle of a statistical business register is that it must be possible to recreate the business frame that was used for a particular publication date. Meanwhile, updates keep flowing into the BR and influencing the view of these historical periods. In general, while revised historical frames could be created, the original frames are used instead. This would be an appropriate strategy for the other base registers. It is information from the time stamped versions of the Statistical Register that we suggest should be used to calibrate estimates and make the results from using "it-is-what-it-is" data sets more representative (less biased) of the target populations. We will discuss this further in Section 5.

In a base register system, there may be a third phase, a feed-back loop between the estimation processes from the evolutionary databases and the register processing. For example, in the context of the BR, business survey responses can lead to updates of name, address, industry classification, and so on. In the case of our evolutionary data bases environment, the estimates that come from the various integrated data sources are analyzed and the knowledge discovered can be fed back to the entity register. These feedback loops are important for keeping the core registers up-to-date and as accurate as possible. We assume the updates will be tombstone information at the entity level.

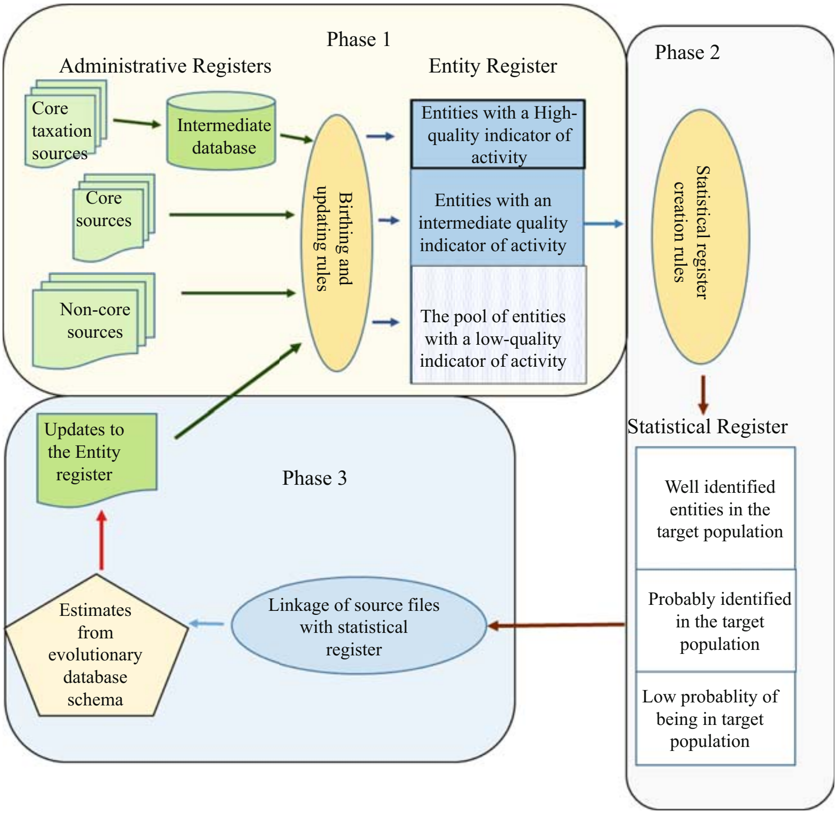
Phase 1
Administrative Registers
Entity Register
Core taxation sources
Intermediate database
Core sources
Non-core sources
Birthing and updating rules
Entities with a High-quality indicator of activity
Entities with an intermediate quality indicator of activity
The pool of entities with a low-quality indicator of activity
Phase 2
Statistical register creation rules
Statistical Register
Well identified entities in the target population
Probably identified in the target population
Low probablity of being in target population
Phase 3
Updates to the Entity register
Estimates from evolutionary database schema
Linkage of source files with statistical register

Fig. 4. Putting it all together: a base register system.

Event information is not maintained in the register. Observed shortcomings in the register's design may also lead to feedback. This can be particularly important in a build-up stage to monitor and tune the rules for "birthing" and "statistical register creation" in phases 1 and 2. As the CSAs become more proficient at generating these intermediate files, they will provide the registers with a continuous maintenance function. Figure 4 illustrates all three phases as one integrated register maintenance system.

4. The Evolutionary Database Schema – The Data Model

4.1. Time and Cause and Effect Relationship

CSAs tend to view time as a descriptive characteristic rather than a fundamental dimension. Time becomes an estimation domain much like sex, age, race, and so on. Yet, observations are events in time and when we combine two or more data sources we need to know how to time order the events observed in the data sources. Many social scientists intuitively grasp this point because they are looking for cause and effect relationships or they wish to understand how social systems are evolving. For social scientists, time is a transcendental variable that helps them make sense of estimates. CSAs tend to think in terms of cross-sectional estimates (or panels) in time rather than a time series evolution. Time series analysis questions are often “end of the line” analyses that marginally affect the cross-sectional survey designs.

Time opens avenues for us to use, analyze and improve the quality of our data sets. Observing related events (a timeline of events) for an entity can provide us with a sense of the evolutionary changes in our data or the volatility of measurements over time. This can provide us with proxies for measuring quality. Having a timeline of events for a common individual allows us to develop improved methods for detecting and fixing errors that are localized to one time period. When one wishes to link entities and events in disparate data sets, the time lines can provide extra information that can improve the quality of the linkages and in some cases, it may allow us to develop quality measurement tools for the cross-linkages.

Time in our evolutionary schema is a fundamental concept and every recorded event must have a time stamp. There is a time-ordering of all events in the schema, so we can distinguish between events, such as diagnosis, treatments and results. Time can open new avenues to improve editing, linking and quality measurements. For an interesting and expanded discussion of the importance of time in statistical analysis one might read Dunn (1946).

4.2. Event and Timeline Databases

To illustrate how the evolutionary time schema might work, let us consider an example of a researcher who wishes to test whether a causal link exists between wealth later in life and education. The data sources available are two administrative data sets, an annual filing of income tax returns, and a collection of school records from a group of school boards. We can view each of these data sets as a list of unique entities and associate a set of date-stamped events with each entity. Each entity’s record can be viewed as a timeline of observed events for that entity. Because new events will be constantly added to the database timelines, the timelines are always evolving in time. The events are containers holding the information gathered for this event. In practice, the information might be just a date stamp and virtual pointers to a record in a subsidiary database.

4.2.1. The Annual Filing of Personal Income Tax Returns

Let us call each collection of files that come from a common generating mechanism and contain common frame entities and identifiers a “timeline database”. Thus, the collective of all the taxation filings by individuals through time would be the timelines database of an individual’s annual tax filings. The data within this database would be structured in a specific manner. The fundamental unique key in the database might be the individual’s taxation number and an individual’s annual tax filing would be an event in that individual’s timeline. Note that there is no requirement that the data be collected on a fixed periodicity and entities’ annual filings could be missing in some years. Within each timeline database, the collection of entities must be of a common type, but different timeline databases could contain different types of entities or events. Using survey terminology, there must be a common frame unit within each timeline database. In our schema, the grouping of events based upon a common entity ID within a single timeline

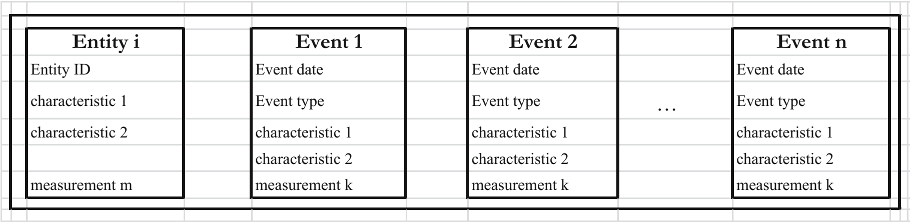

Fig. 5. One entity's timeline in the taxation timeline database.

database is viewed as a deterministic function and not a linking process which we view as a nondeterministic process. Entity IDs are assumed to be known true facts. In our terminology the process of creating the entity timeline in Figure 5 will be referred to as a "grouping" function.

Maintaining the taxation database could be straight forward and cost effective. Whenever a new batch of annual filing comes in, one only needs to find the associated Entity ID in the database and add a new event to the record. If the evolutionary database consisted of containers with virtual pointers, one would only need to update the pointers. If the database was structured properly, this activity might require minimal re-indexing and sorting. Once a new batch of records was appended to the end of the timelines database, it might never be touched again. The imposition of the timelines schema onto the taxation database does not require the owners and previous users of the subsidiary tax databases to change any of their previous methods. If these databases remain static, no changes will occur in the timeline database. Even edits of the subsidiary data sets may not require any updates to the timeline database. Only additions, deletions or changes involving Entity IDs will require a recompilation of the groupings inside the event containers. For an expanded discussion on the evolutionary data base structure, one might read Chapter 3 in Lothian et al. (2017).

4.2.2. The Collection of Education Events from a School Board

For the education timeline database, the unique identifier might be a student while, the events might be results from tests, special education evaluations, discipline reports, and so on. Each event would occur at a specific time and there would be characteristics defined for each event. Some characteristics might be defined at the identifier level (like birth date, last address, name), while others might be defined at the event level (like date of transaction, observed attribute, and so on). In this timeline database, multiple different types of events might be recorded; each being pulled from a different subsidiary database. Again, the records must all relate to a common entity and there must be a mechanism for grouping a student's events into a timeline.

Identifier inconsistencies could be a significant issue with educational data. Changing schools could lead to the generation of an alternate student number and home address. In addition, as children age they can change their desired names. These types of inconsistencies might result in students being assigned multiple identifiers and causing fragmentation of the student's event timelines.

The database design and programming would be evolutionary, so that mechanisms can be developed at future dates to resolve these inconsistencies. The long-term solution for this issue is to make the unique key an internally generated database ID that can be remapped to join up the fragments. For an expanded discussion on this issue, one might read Appendix B in Lothian et al. (2017).

The two timeline databases examples were chosen to illustrate the proposed timeline database structure and give a flavor of the challenges that CSAs would face when implementing this structure. In our design, we emphasize flexibility and evolution to deal with the constant changes in the number of data sources and what is contained within each source. By "farming out" the control of the data sources we are implicitly accepting that each source is an "it-is-what-it-is" data source. Note that timeline databases are not lighthouse registers, but the lighthouse registers might use information from timeline databases.

4.3. Linking or Relating Timeline Databases

One of the intended purposes of the evolutionary schema is to allow users to explore cause and effect relationships and to cross-relate disparate "it-what-it-is" data sets. To accomplish this task, users of this evolutionary schema will want to build and discover linkage relationships between different timeline databases. As an example, they may want to explore the relationship between school performance and health and wealth in three data sources. To build the output data set, one must start with one of the timeline databases and connect to a second timeline database through a linkage relation database. Then one must connect the resultant amalgamated database to the third source database using a second linkage relation database. Figure 6 illustrates how the potential relationships (linkages)

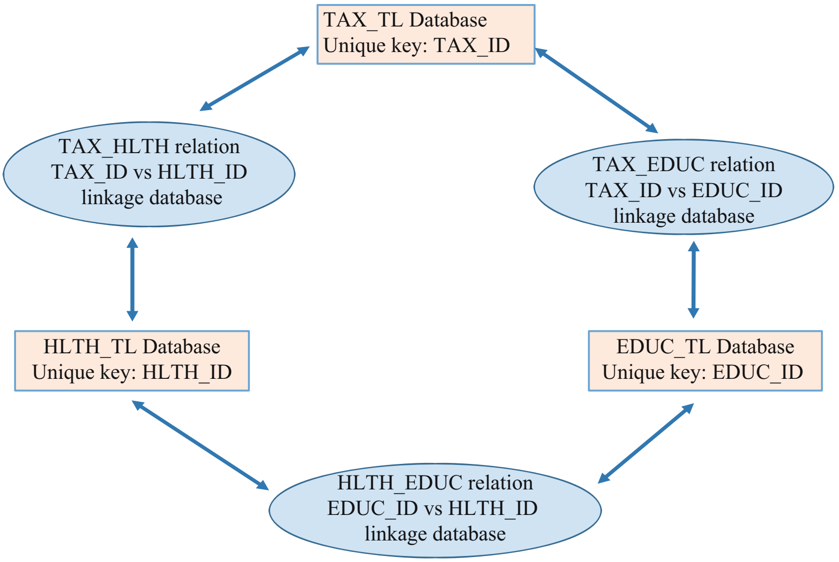

Fig. 6. Functional relationships in the evolutionary system.

between the timeline databases are defined. The ovals in Figure 6 will be referred interchangeably as linkage functions or linkage databases.

The linkage databases are the tools that will allow us to cross-relate timeline databases and explore causal models. The linkage databases will contain unique key pairs defining relationships (links) between two timeline databases. All linkages are assumed to be one-to-one and are not necessarily exhaustive. Only direct relationships can exist between two timeline databases. Linkages involving three or more timeline databases only exist indirectly and the solutions are path dependent. Thus, if A, B, and C are three timeline databases then $(A \cap B) \cap C \neq (A \cap C) \cap B$. The intersection of the three databases is path dependent and is neither commutative nor transitive. Anyone who has observed Google searches is aware that the results of the search depend on the order of the words used to do the search. This can lead to inconsistencies in the produced results, but a generalized multi-path linkage function is not feasible. CSAs will have to establish orders of precedence for multi-source linkages.

Our database design is evolutionary in many senses. The algorithms linking, editing or transforming the data will incrementally evolve as more knowledge is acquired. Initially, some linkage relationships might be undefined and implemented on a need-to-have basis. Fields in subsidiary databases, events, timelines and new survey data sources can be iteratively added as they become available or are needed. The maintenance of the evolutionary database will be devolved and distributed amongst local groups with strengths and experience in the local data. The timeline databases can be disseminated among unrelated control groups and separate teams could be assigned responsibility for creating and maintaining the linkage databases. Each team could add events; edit fields independent of the other groups. New linkage technologies could be implemented without requiring any revisions to the timeline databases or their subsidiary databases. It becomes a distributed and cooperative evolutionary system where local changes will not force a recompile of the complete system.

4.4. Cross-Linking Disparate Data Sets Significantly Increases Risks

Intuitively, most humans have a sense of the Law of Large Numbers. We know that small observation sets are untrustworthy. So, it is natural to assume that adding more data to our system must make it more trustworthy, but that is not true when cross-linking timeline data sources. For linked data sets, this premise does not hold. To demonstrate the problem, let us return to the example we used in Figure 6, where we are linking health and wealth data sources, so we can explore how health affects wealth. Let us make a few conceivable assumptions. We will assume that each data set is drawn for the same population, but suffers from coverage issues. Perhaps, the TAX data does not cover some of the people who never entered the labor market and thus earns no revenue and the health data only partially covers children. Figure 7 illustrates what happens when you link the data. The integrated data set (small oval) acquires the weaknesses of both source data sets. Integrating additional data set sources only makes things worse. When dealing with integrated data sets, one should always assume the output linked data set will not be representative of the target population.

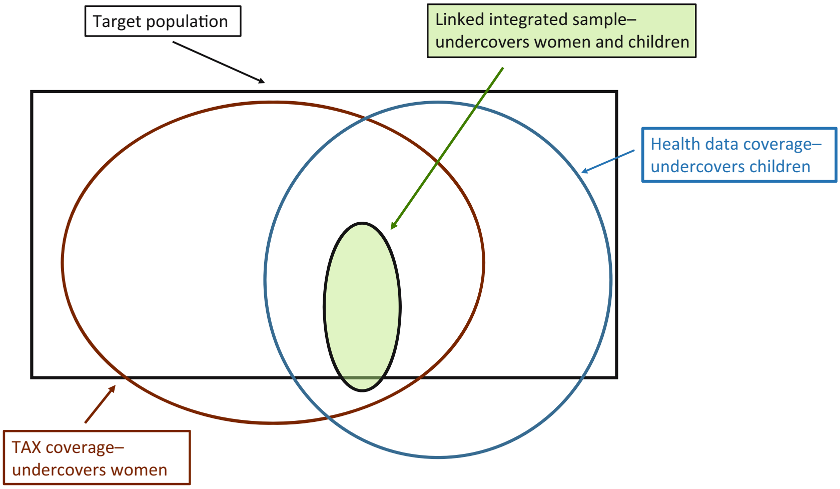

Fig. 7. Coverage biases in integrated data sets.

Representativeness is the central issue when dealing with it-is-what-it-is data sets and this section illustrates that however we attack the problem, there are no magic bullets. Instead, with representativeness in mind, one must use a structured and methodical strategy to identify and eliminate biases related to coverage problems. Our approach transfers the representative problem to the lighthouse registers in Figure 1. There dedicated teams can focus on continually identifying and improving representativeness in an evolutionary manner. Like in the case of the BR, new sources or strategies will be found to minimize coverage biases.

4.5. *Timelines are a Fundamental Concept*

In the data model presented, time is a foundational concept. Our intent was to design a database that could relate causes and effects, and a necessary requirement for this is ordering events from multiple timeline databases into a single event timeline. The linking of events into a timeline can open avenues for improved linkage strategies. Missing linkage variables can be estimated from other events in the timeline and inconsistencies in names, addresses, age, and so on, can be edited and standardized by analyzing the full timeline. The linkage strategy could depend on time vectors instead of a single value. Perhaps this kind of linkage strategy could help us deal more effectively with name and address changes.

4.6. *Linkage Processes have an Underlying Probabilistic Nature*

Linkage functions are assumed to be probabilistic, in the sense that the linkage function always has significant uncertainty associated with each identified link. Linkage functions produce a subsample of the two source data sets, where the records produced have an underlying probabilistic element. The number of links found and the "truth" of each link is probabilistic (random) in some sense. While it is almost certainly true that the linked data set is generated by some probabilistic process, we have little knowledge concerning the selection probabilities. We are not even certain whether the selection is with or without replacement. A key issue is whether the probabilistic sample is representative of our target population. What we will assume is that the sample selected will be non-confounded (at random) (Rancourt et al. 1994) within some estimation domain. This is a powerful assumption.

Our linked output data set is a convenience sample (i.e. non-probabilistic samples sometimes referred to as opportunity or accidental samples see for example Baker et al. 2013) which we will refer to as a linked sample. Our linked sample is a subpopulation derived using relationships to which we have access, but we have no control or knowledge of how these relationships were constructed. Researchers using this linked sample cannot make scientific generalizations about the general population from this sample because it may not be representative of the target population. Strictly speaking, linked or convenience samples are non-probability samples (Baker et al. 2010 and 2013), yet we can hypothesize a hidden underlying probabilistic selection mechanism that is random within some estimation domain (stratum). Thus, the non-probabilistic element of the selection process only affects the balance between estimation domains. The credibility of a researcher's results when using this hypothesis will depend on convincing the reader that the researcher has properly compensated for the imbalance between domains and that the final estimates are representative of the population of interest.

4.7. Using Stratum Definitions to Improve Representativeness

CSAs create various types of strata to help address nonrepresentative issues and to improve the efficiency of estimates. Coverage issues are regularly encountered in CSA surveys and censuses. Even full-enumeration censuses can experience significant undercoverage of special subpopulations. CSAs have several strategies for dealing with these types of issues. One standard practice is to assume that nonresponse is missing at random within a stratum. This is analogous to the strategy that we are proposing.

CSA methodologists are mindful of the potential weaknesses caused by assuming that nonresponse is missing at random. Yet, nonresponse and undercoverage of special subpopulations occurs in every survey and census. This has forced CSA methodologists to develop a toolbox of strategies to eliminate coverage issues. These may include: assuming missing at random within strata; modelling using auxiliary information; Bayesian imputation; targeted follow-up surveys of under-represented subpopulations; calibration; capture-recapture techniques; propensity models; and so on.

Historically, CSAs have followed an evolutionary strategy in developing these methodologies. The authors see a comparable evolutionary development strategy occurring within our schema. Our assumption of "random selection within a stratum" is a first step in this development chain. We expect that, over time, more sophisticated technologies for dealing with coverage issues will arise.

Our lighthouse registers and assumption concerning randomness within a stratum are initial building blocks that force methodologists to confront representative issues head on. What we are proposing is a heuristic strategy based on what worked in the past, rather than a theory built from first principles. If a CSA can construct the three lighthouses and if the statistical registers are a reasonable approximation of complete unduplicated coverage of the target populations, then the authors believe that toolboxes for fixing representativeness can be developed for "it-is-what-it-is" data sets.

5. Estimation Plays a Central Role in the Design of the Evolutionary Schema

Up to this point, the focus of our database design was implementing a cause and effect design and exploiting the scalability, flexibility and efficiencies of an evolutionary data model. The objective is to use estimates derived from this schema to make inferences concerning real world populations and how entities in these populations interact in time and space. (For a further discussion on the importance of space and supporting GIS solutions in our paradigm see Lothian et al. 2017).

To derive interpretable estimates from the data, we must make some conjectures about the data and apply a structured scientific estimation theory. From survey theory, we suggest borrowing from one of the three different paradigms. The first is the design-based, or randomization, theory (Särndal et al. 1992), which emphasizes that attribute values of the records in the data are fixed values and that it is the random selection of the elements in the data set that ensures the representativeness of the target population through the use of a sampling frame. The second paradigm is the predictive, or model-based, approach (Valliant et al. 2000), where the values are regarded as realizations of random variables and the design by which the survey elements are selected is of less importance. A third paradigm that can be implemented is a Bayesian inference framework. It has been put forward as useful for analysis of small non-probabilistic samples and appears to be a direct alternative when data from different sources are being combined (Little 2012 and 2015; Rao 2011). The choice of which paradigm to use depends both on one's estimation objective and/or philosophical training. In our schema, we present a design-based paradigm, but either of the other two paradigms could be substituted. Our schema does not favor any of the three paradigms. One just needs to choose one paradigm and stick with it.

In the literature to date, there has been much discussion on the data availability, potential data models, building the databases and linking algorithms, rather than how estimation and statistical inference enters the data schema. Most of the literature seems to focus on database structures or building the information technology infrastructure (Holman et al. 2008). There are numerous articles on linkage algorithms (Fellegi and Sunter 1969; Jabine and Scheuren 1985; Winkler 2009); others on specific attempts to define the required data structures (Holmberg et al. 2011; Wallgren and Wallgren 2014); and others on the construction of specific data ecosystems (Holman et al. 1999). Literature with an end-to-end perspective of all the components necessary to do statistical inference and estimation are less common. One such overview is the paper by Zhang (Zhang 2012), which provides a conceptual statistical methodological framework for using "it-is-what-it-is" data sets (or administrative data sets in his article). This article was inspired by Zhang's article. Our schema provides one possible implementation framework within Zhang's conceptual model.

Discussions on estimation strategies are complex and the non-probabilistic nature of the "it-is-what-it-is" data sources can generate considerable controversy. As we mentioned previously, the major obstacle is constructing an estimation framework that generates results that are representative of the general population. We believe that registers must play a central role in making estimation representative. Registers and frames are the support scaffolding for estimation done within the evolutionary schema. Without this scaffolding, our estimation strategies will be weak and prone to failure irrespective of the design paradigm that we choose to use.

5.1. Registers/Frames Anchor the Evolutionary Databases

Generally, statistical frames will be derived from information available from one of the three base statistical entity registers using the Statistical Register Creation Rules (Figure 3). If U^R is the set of all entities in our base register, then:

$$U^F \subset U^R \tag{1}$$

The base or entity register U^R is an integrated, micro-merged and maintained list of entities created from the combination of different administrative data sources based on identifiers that are unique to the various data sources, (see Section 3). In addition, we need a linkage relationship database that cross-links the statistical frame and the entity timeline database of interest. With U^F defined, we could calibrate the "it-is-what-it-is" linked sample to the frame. (In Bayesian terminology, these calibrations would be prior constraints on the probability distributions.) By using aggregate data from the frame, such as domain totals X_d and domain counts N_d and regarding them as known constants, weights and calibration equations can be constructed that reproduce these known parameters within the linked data set. Hence, we construct weights w_k that satisfy the principal expressions:

$$\begin{aligned} X_d &= \sum_{A_d} \omega_{d,k} x_{d,k} = \sum_{U_d^F} x_{d,k} \\ N_d &= \sum_{A_d} \omega_{d,k} = \sum_{U_d^F} 1_{d,k} \end{aligned} \tag{2}$$

where A_d is the subset of linked elements k observed within domain (stratum) d. U_d^F is the complete enumeration of elements k within domain d in the survey frame F while $\mathbf{x}_{d,k}$ is a vector of known variables provided by the frame. Depending on the circumstances, different variants of Equation (2) can be used (Särndal et al. 1992).

In our formalism, we assume that U^F provides unduplicated complete coverage of the target population U^T. To cover domains d, we just add the subscript d and define variables

$$x_{d,k} = \begin{cases} x_k & for \quad k \in U_d^F \\ 0 & for \quad k \in U^F - U_d^F \end{cases} \tag{3}$$

where an important case of x_k are the indicators 1_k variables that gives us, $N_d = \sum_{U_d^F} 1_{d,k}$. The estimation of other parameters, for example, an unknown total $Y_d = \sum_{U_d^F} y_{d,k}$ from the linked data set is then done by using the same set of calibration weights, that is, $\hat{Y}_d = \sum_{A_d} \omega_{d,k} y_{d,k}$.

This is a popular technique often used in a design-based inference (Lundström and Särndal 2005; Särndal 2007). To apply it here, we have to make the naïve assumption that the linked sample within an estimation domain is a nonconfounded sample from the target population. Then calibrated estimates are possible, and under this simple scenario one could estimate variances. (Nonconfounded might be considered as synonymous with "observations missing at random". We assume that the observed subpopulation is representative of the full target population in every variable. Nonconfounded has a wider contextual meaning that implies the measurement variable is unbiased in both the statistical and non-statistical sense. The term "unconfouded" is discussed by Rancourt et al. 1994).

If we have entity level information from the frame, we can take the calibration technique one step further and apply explicit models, that is, compute calibration weights that, instead of Y_d and N_d produce model predictions of super-population parameters Y_d^* and N_d^* (Wu and Sitter 2001). In this case, a separate model can be applied for every study variable and domain, although that would require a considerable modelling effort.

In a Bayesian framework, we can denote the information we have from the statistical entity frame by Z, it would enter in the specification of the prior distribution of the population values, $p(Y|Z)$. It would also be used as covariates during the generation of the posterior distribution and parameter estimation when the prior is confronted with the linked data (Bryant and Graham 2015).

The above discussion is a framework for an estimation strategy rather than an explicit methodology. In practice, we are advocating reproducing classical (design-, model-, or Bayesian-based) estimation techniques used in current survey designs.

5.2. *Linking Disparate Timeline Source Files to the Frame*

Ancillary register/frame information is necessary for unbiased estimation when one is cross-linking two "it-is-what-it-is" data sets. To illustrate this point, we will demonstrate how estimation might occur in a simple example. Let us look at the Section 4 example with our timeline databases of health and income. We wish to estimate the relationships between health and wealth of individual entities in the current population (see Figure 7).

There are multiple difficulties with these data sets. First, the TAX and HLTH timeline databases contain out-of-scope units and have significant undercoverage of the target population. In survey terminology terms, we are neither sure of the true target population size N nor of the true sample size n. Without some knowledge of the true n and N, how can we make unbiased estimates of the relationships? How do we answer such questions as: how many children are expected to have a specific health issue? At best, we can estimate ratio or proportional effects that apply to unknown subpopulations. Second, we know that both timeline databases are confounded, possibly in different ways. In general, the poor, immigrants, the very young, the very old, persons with handicaps, stay-at-home parents, and so on may be missing from one or both timeline databases. In survey language terminology, our linked subsample is a biased sample. However, if we assume that the sample within each estimation domain d is "at random" or nonconfounded, and if the entity frame is of good quality, giving us n_d and N_d we can apply the calibration technique in Equation (2) and improve the estimates by decreasing the bias. Without the register/frame we would not have this possibility.

5.2.1. Linking the Integrated Sample to the Frame

There are two simple strategies one might use to link the integrated sample to the empirical frame. If we are fortunate enough to have quality information available to link the sampled records at the entity level, we can create a micro-entity estimation file with weights applied at the entity level. Even if entity level linkage is not possible, we could link at the domain estimate levels if each timeline database file contains the domain stratum identification variable. In these simple cases, we will be making the naïve assumption that any over- or undercoverage or nonresponse is "at-random" and within an estimation domain is ignorable. As mentioned in Subsection 4.7, CSA methodologists have developed a toolbox of strategies to eliminate coverage issues. We are presenting one of the most straightforward strategies.

5.2.2. Estimation when Entity Level Linkage with the Frame Exists

Let us assume that we have a statistical register of current residents (perhaps derived from a recent census or as a result of maintaining a PR lighthouse). Furthermore, a linkage relationship exists, connecting the resident ID on the frame to the person's personal TAX_ID in the taxation timeline database.

Each ellipsoid in Figure 8 could be considered indicative of a survey processing step or a linkage-step and we call the steps through the linkages 'phases'. Figure 8 illustrates what a phase means in our schema: the pentagon is our output integrated data set from the first phase and comprises two container fields holding the TAX_ID and the HLTH_ID. Without context, the output data set is not generally sufficient information to derive reasonable estimates. In phase 2, the context (or calibration) turns the phase 1 output file into estimates. At times, the second contextual phase can be overlooked when dealing with "it-is-what-it-is data". We believe this is a key point, and frames and registers can provide the scaffolding that supports quality estimates.

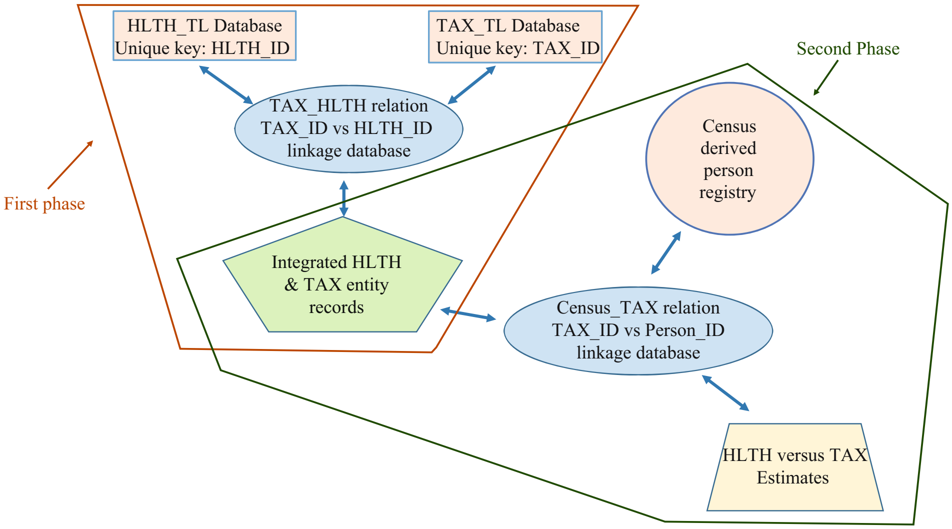

Fig. 8. Estimation processing steps when entity level linkages exist.

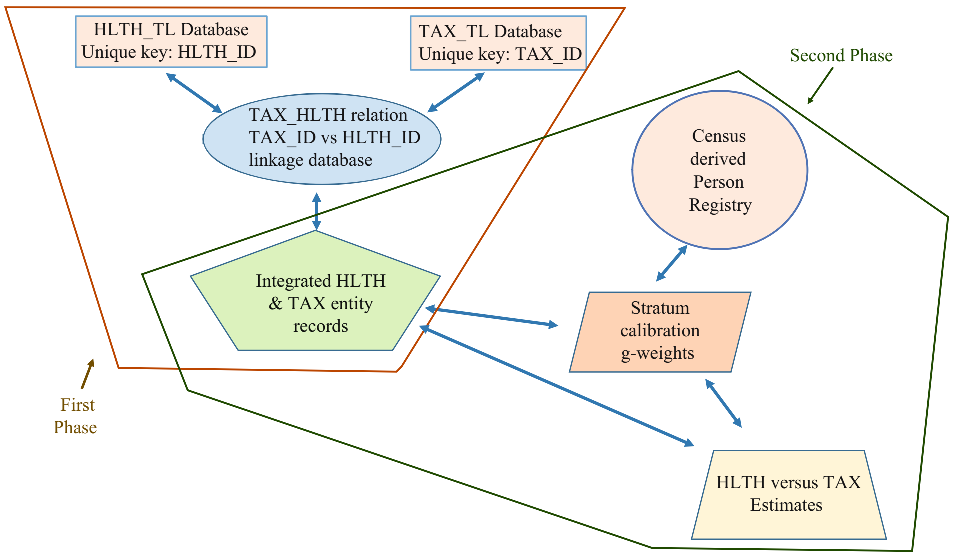

Fig. 9. Estimation processing steps when only common stratum information is available.

5.2.3. Estimation when Entity Level Linkage with the Frame Does Not Exist

In some cases, no reliable entity level linkage information may be available, or it is possible that we wish to link disparate information collected from different, but similar, entities. If the three data sets (health, wealth, census frame) are nonconfounded within an estimation domain, are concurrent and have common domain stratification variables, then an estimator may be found. Figure 9 illustrates this case.

In Figure 9, the linkage between the entity level data (pentagon) and the frame (the circle) will be indirect. The second phase parallelogram represents the combined information from the two. As an example, assume the linked timeline databases contain categorical data such as a geographical (domain d with D categories) and a socio-economic classification (domain d' with D' categories) and that we have this information in the frame as well, but not necessarily simultaneously. With ς being the domain category indicator, we can form $\mathbf{x}_k = (\varsigma_{1k}, \ldots, \varsigma_{dk}, \ldots, \varsigma_{Dk}, \varsigma_{1'k}, \ldots, \varsigma_{d'k}, \ldots, \varsigma_{D'k})^T$ and the right-hand side of (2) will be the domain counts $N_{d\bullet}$ for $d = 1, \ldots, D$ and $N_{\bullet d'}$ for $d' = 1, \ldots, D'$. Hence, we use the marginal distribution of the domain categories from the frame to support the estimation. The weights that satisfy a calibration equation will depend on the inferential principle used and with this $\mathbf{x}_k$ there is no nice expression, however numerical computations are not difficult. In a generic sense, Equation (2) still holds.

6. Total Survey Error Measurement in the Evolutionary Schema

6.1. Error Measurement and Metadata

The authors see information on error measurement as a fundamental component of the evolutionary schema. We see this information being stored and maintained in ancillary metadata files within the ecosystem. We see a metadata file attached to every register, source data and linkage function file in the ecosystem. The metadata file will contain information on data sources, variables available, discussions on weaknesses and strength of the data and other quality-related information. We see these files as a vital component of ensuring representativeness and long-term quality of estimates.

6.2. Metadata is the Gateway to the Evolutionary Schema

Researchers will often approach the evolutionary database with an ambiguous research objective. While these research questions may appear to be wide-ranging or imprecise, they often have very restrictive underlying constraints that will impact on estimation, such as, requiring specific data years and/or subpopulations and/or source data sets and/or relationships being estimated. Researchers will want to know if the available sources/relationships/registers/linkage functions adhere to these constraints. Thus, associated with every object/function in the evolutionary schema should be a metadata descriptor.

The most basic and most requested metadata is an explanation of how to access the data and how it is structured. Users wish to get on with using the data. They require database access protocols, file names and locations, field names, formats and brief descriptions, and source providers for the various data sources.

While this access information is vitally important to users, it presents dangers if it is not balanced with information related to the quality and limitations of the data. Without some discussion on the populations covered by each file, you are encouraging users to apply their analysis to inappropriate subpopulations. Reid et al. (2017) propose a framework for quality documentation and communication in this situation. Or, perhaps, the user is linking two sources with slightly different entities (family versus head of household), and this creates misleading relationships. The authors believe that the priority of meta-data should be the presentation of quality information to the database user, rather than a focus on metadata that is easy to create or requested most often.

6.3. Measuring TSE

Total Survey Error (TSE) (see, for example Biemer 2010; Groves and Lyberg 2011) can rarely be reduced to one number, instead it is a structured methodological approach for reviewing and compiling sources of error in a survey. Errors can arise at each classical survey processing step: frame creation, sample design, questionnaire design, questionnaire distribution, collection, editing, follow-up, imputation and estimation. Each survey processing step can introduce errors in potentially different dimensions of error. The TSE paradigm treats each aspect or dimension of the survey processing system as part of a collage that defines the overall measure of quality. At each processing step, quality measurables are collected and assembled into an overall package that allows the survey designer to understand where errors occur and give them some sense of their overall impact on the TSE. TSE is a structured approach to cataloguing and measuring the errors that arise in each of the classical survey processing steps. TSE challenges us to view survey errors in a structured and holistic manner.

While the classical survey processing steps may not be applicable in a data integration paradigm, we believe one should use a similar structured approach that breaks down the overall estimation process into self-contained subprocesses. A few simple measurables will be suggested within each subprocess. These measures will be naïve, but we believe they will address the primary concerns of most users. We propose focusing on the coverage or representativeness of the specific output data set in the specific subprocess.

6.4. *Error Arising from Source or "it-is-what-it-is" Data Sets*

In our schema, the source data sets are the rectangles in Figures 8 and 9. Source data sets will be administrative, survey or census data sets containing observed variables that we wish to interrelate. Often, the incoming quality and content of these sources will be beyond the control of the CSAs. In the following section, quality measurables will be suggested for these source data sets. Note that the timeline concept introduces a new way to view quality in the schema.

6.4.1. Coverage Statistics

In the Evolutionary Schema, population coverage is a critical concept because it is expected that most data sources will have biases in their coverage of the target population. Every source data set should be related to its coverage of a frame or register, preferably either the BR or the PR or the land register. The population coverage should be as devolved as possible. Thus, for person entity data sources one should provide coverage by sex, age, geography and as many other demographic characteristics as possible.

Thus, for each processing step p and domain d, we will calculate the quality measure set $Q_d^{p,R} = \langle N_d, n_d, f_d \rangle_{p,R}$ where N_d is the size of the target population in domain d of frame U^F which in turn was derived from the base register U^R. Let n_d will be the number of entities in domain d observed in the output data set. Finally, coverage for process p within domain d will be

$$f_d = \frac{n_d}{N_d} \tag{4}$$

If there are m domains in processing step p, the full set of preliminary quality measures $Q_d^{p,R}$ will be:

$$Q^{p,R} = \left\{ Q_{d_1}^{p,R}, Q_{d_2}^{p,R}, Q_{d_3,\bullet,\bullet,\bullet,}^{p,R} Q_{d_m}^{p,R} \right\} \tag{5}$$

By examining $Q^{p,R}$ for all d, the analyst/user can derive some sense of the representativeness of the output data set derived from process p. These three measures are simple, and far from comprehensive, yet nevertheless powerful. They address the principle weakness of linked and "it-is-what-it-is" data sets, representativeness. Small values of n_d and f_d or disparate values of f_d for different d can be indicators of quality problems.

We propose using at least one of the three statistical registers/frames to generate $Q^{p,R}$ for each processing step and then placing this information in the metadata. We recognize that these sets of measures are not comprehensive, but they are relatively simple to autogenerate and if they are used properly they will stop the worst cases of misuse of the data. From our perspective, this is the first step in the evolutionary development of TSE indicators of representativeness.

6.4.2. Stability Over Time

These basic measures could be augmented by measuring changes within event timelines. Timelines are collections of events for a specific entity from a common data source. For each entity's timeline, we could record statistics concerning how often a field changes or is missing. Then, aggregate (domain) estimates of the average changes or proportions of missing values could be automatically generated and placed in the metadata. Grouping events into timelines gives us an extra dimension of quality to measure.

6.4.3. Evolution will Develop New Measures of Quality

As new registers, sources, linkage functions, timelines and events get added to the database, users will develop new insights about the data and better measures of quality. We will discover new algorithms to calculate these measures and place them into the metadata.

7. Summary

We set out to identify some key issues in using administrative data: estimation and assurances of quality. In a two-year-long discussion amongst ourselves and other methodologists, we explored the numerous pitfalls one encounters when using administrative data and we discussed several strategies that needed to be a part of any statistical system using administrative data. While we quickly realized that representativeness was the Achilles heel of administrative data, we were strongly influenced by Zhang's article calling for a new conceptual paradigm when dealing with administrative data. Thus, right from the beginning of our discussions we attempted to tackle the problem in a holistic manner, attempting to use a full conceptual paradigm for dealing with administrative and it-is-what-it-is data. Below is a summary of our major finding and an outline of the main features of our schema.

7.1. Administrative Data is Nonrepresentative

The key weakness of administrative data is that various sections of the targeted population have coverage issues, and this generates representativeness problems. Coverage is never 100% in any administrative data source and in most cases, significant portions of a population are under- or overcovered. This is a systemic problem that is considerably worsened by cross-linking multiple administrative data sets. Methodologists must address this key fact steadfastly. We must be capable of defending our estimates from criticisms of bias caused by undercovering the under-priviledged or the rare populations. If we cannot do this effectively then CSAs will lose credibility.

7.2. Correction with Registers and Frames

Our solution to this problem and recommendations to CSAs is to create the three lighthouse registers and use them to measure under- and overcoverage in various domains/strata/classes. Then we use these registers to create calibrations (design-based designs), models (model-based designs), or Bayesian priors (Bayesian designs) to correct for coverage issues. Here we are following the historical development path for correcting coverage issues in censuses.

7.3. Evolutionary (System Grows in Every Sense Over Time)

We see our schema as an evolutionary system in every sense. New data sources will evolve, and old ones will disappear. New data points will be added (both in time and cross-sections). New database designs will be incorporated, and new estimation and linkage algorithms will constantly be developed. New methodologies will be constantly under development and evaluated.

7.4. Distributed and Collaborative System

Administrative data spans wide areas of knowledge, subject-matter areas, geography and time. In addition, the final design will incorporate ongoing development of complex statistical and IT methodologies. No one group can do these tasks centrally. The tasks and data sources must be delegated across various teams with varying backgrounds and expertise. A central design and control structure would be created to oversee these teams.

7.5. Evolutionary Convergence

When we create our registers, data sources, methodologies, and so on, there must be a path of convergence towards an ever-improving system. Ideally, each new evolutionary step will incorporate all previous information gathered. Consider the BR. In general, most of the information in the BR is tombstone information that rarely changes over time. Births and deaths are a small percentage of the population, only a small percent of the addresses or names change each period, and so on. The BR team focuses on changes rather than the full population. This is also the manner in which the census address list is maintained. Similarly, our evolutionary system would be built along paths that evolve towards better quality and optimality.

Feedback loops are important in this system. Users must be able to feed corrections that they have identified in the registers, algorithms, and so on, back into the system. This is the way the BR and the address register of traditional censuses worked.

7.6. Timelines (Cause and Effect)

There is one message we are hearing continually from researchers who want to use CSA data. They want to do cause and effect studies or longitudinal studies. Our data need to have a natural time structure built in from the beginning. We see each administrative data transaction as an event that occurs at a specific time. We propose grouping and time ordering these events that occurred for a common entity into timelines. For each data source, an implicit timeline database would be created for each entity. Viewing the data in this manner not only allows for effective studies, it also opens new possibilities for editing, imputation, linking, and so on. Of course, there are still statistical challenges in establishing representativeness in longitudinally linked records in order to reliably interpret the results.

7.7. Total Survey Error

We propose a simple strategy for creating measurement tools for quality estimation and we address the representativeness side of the Total Survey Error (TSE) model. In each stratum/class/domain defined by the three lighthouses, we propose calculating the coverage ratio of that data source versus the estimated lighthouse population. When cross-linkages and integrations are done, the stratum coverage ratios should be calculated for the linked data set. This TSE information should be stored in the metadata.

8. References

Baker, R., S.J. Blumberg, J.M. Brick, M.P. Couper, M. Courtright, M. Dennis, D. Dillman, M.R. Frankel, P. Garland, R.M. Groves, C. Kennedy, J. Krosnick, P.J. Lavrakas, S. Lee, M. Link, L. Piekarski, K. Rao, R.K. Thomas, and D. Zahs. 2010. “AAPOR Report on Online Panels.” *Public Opinion Quarterly* 74(4): 711–781. Doi: https://doi.org/10.1093/poq/nfq048 (accessed May 2018).

Baker, R., J.M. Brick, N.A. Bates, M.P. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, and R. Tourangeau. 2013. “Summary Report of the AAPOR Task Force on Non-Probability Sampling.” *Journal of Survey Statistics and Methodology* 1(2): 90–143. Doi: https://doi.org/10.1093/jssam/smt008 (accessed May 2018).

Bakker, B.F.M. and P.J.H. Daas. 2012. “Methodological Challenges of Register-based Research.” *Statistica Neerlandica* 66(1): 2–7. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00505.x (accessed: May 2018).

Biemer, P.P. 2010. “Total Survey Error: Design, Implementation, and Evalutaion.” *Public Opinion Quarterly* 74(5): 817–848. Doi: http://dx.doi.org/10.1093/poq/nfq058 (accessed May 2018).

Bryant, J.R. and P. Graham. 2015. “A Bayesian Approach to Population Estimation with Administrative Data.” *Journal of Official Statistics* 31(3): 475–487. Doi: http://dx.doi.org/10.1515/JOS-2015-0028 (accessed May 2018).

Dunn, H.L. 1946. “Record Linkage.” *American Journal of Public Health* 36(12): 1412–1416. Doi: http://dx.doi.org/10.2105/AJPH.36.12.1412. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1624512/ (accessed May 2018).

Fellegi, I.P. and A.B. Sunter. 1969. “A Theory for Record Linkage.” *Journal of the American Statistical Association* 64(328): 1183–1210. Doi: http://dx.doi.org/10.1080/01621459.1969.10501049 (accessed May 2018).

Ferrara, A., A. Nikolov, and F. Scharffe. 2011. “Data Linking for the Semantic Web.” *International Journal on Semantic Web & Information Systems* 7(3): 46–76. Doi: http://dx.doi.org/10.4018/jswis.2011070103 (accessed May 2018).

Fowler, M. and P. Sadalage. 2003. Evolutionary Database Design. Available at: http://martinfowler.com/articles/evodb.html (accessed May 2018).

Groves, R.M. and L. Lyberg. 2010. “Total Survey Error: Past, Present, and Future.” *Public Opinion Quarterly* 74(5): 849–879. Doi: http://dx.doi.org/10.1093/poq/nfq065 (accessed May 2018).

Hand, D.J. 2018. “Statistical Challenges of Administrative and Transaction Data.” *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 181(Part 3): 1–24. Doi: http://dx.doi.org/10.1111/rssa.12315 (accessed May 2018).

Holman, C.D., A.J. Bass, D.L. Rosman, M.B. Smith, J.B. Semmens, and F.J. Glasson. 2008. “A Decade of Data Linkage in Western Australia: Strategic Design, Applications and Benefits of the WA Data Linkage System.” *Australian Health Review* 32(4): 766–777. Available at: https://www.ncbi.nlm.nih.gov/pubmed/18980573 (accessed May 2018).

Holman, C.D., A.J. Bass, I.L. Rouse, and M.S.T. Hobbs. 1999. “Population-based Linkage of Health Records in Western Australia: Development of a Health Services Research Linked Database.” *Australian and New Zealand Journal of Public Health* 23(5): 453–459. Available at: https://www.ncbi.nlm.nih.gov/pubmed/10575763 (accessed May 2018).

Holmberg, A., K. Blomqvist, J. Engdahl, H. Irebäck, L.-G. Lundell, and J. Svensson. 2011. *A Strategy to Improve the Register System to Store, Share and Access Data and its Connections to a Generic Statistical Information Model (GSIM)*. Paper presented at the Work Session on Statistical Data Editing, UNECE, Ljubljana, Slovenia, May 9–11. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2011/wp. 37.e.pdf (accessed May 2018).

Holt, T. 2000. “The Future for Official Statistics.” *Journal of the Operational Research Society* 51(9): 1010–1019. Doi: http://dx.doi.org/10.1057/palgrave.jors.2600999. Available at: http://www.jstor.org/stable/254222 (accessed May 2018).

Jabine, T.B. and F.J. Scheuren. 1985. “Goals for Statistical Uses of Administrative Records: The Next 10 Years.” *Journal of Business & Economic Statistics* 3(4): 380–391. Doi: http://dx.doi.org/10.2307/1391725 (accessed May 2018).

Kruskal, W. and F. Mosteller. 1979. “Representative Sampling, II: Scientific Literature, Excluding Statistics.” *International Statistical Review/Revue Internationale de Statistique* 47(2): 111–127. Doi: http://dx.doi.org/10.2307/1402564. Available at: http://www.jstor.org/stable/1402564 (accessed May 2018).

Langer, G. 2013. “Comment: Summary Report Of The AAPOR Task Force On Non-Probability Sampling.” *Journal of Survey Statistics and Methodology* 1: 130–136. Doi: http://dx.doi.org/10.1093/jssam/smt008 (accessed May 2018).

Little, R.J.A. 2012. “Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics.” *Journal of Official Statistics* 28(3): 309–334. Available at: http://www.jos.nu/Articles/abstract.asp?article=283309 (accessed May 2018).

Little, R.J. 2015. “Calibrated Bayes, an Inferential Paradigm for Official Statistics in the Era of Big Data.” *Statistical Journal of the IAOS* 31: 555–563. Doi: http://dx.doi.org/10.3233/SJI-150944 (accessed May 2018).

Lohr, S.L., V. Hsu, and J.M. Montaquila. 2015. *Using Classification and Regression Trees to Model Survey Nonresponse*. Paper presented at the Joint Statistical Meeting (Section on Survey Research Methods), Seattle, Washington, United States. Available at: https://ww2.amstat.org/sections/srms/Proceedings/y2015/files/234054.pdf (accessed May 2018).

Lothian, J., A. Holmberg, and A. Seyb. 2017. Linking Administrative Data: An Evolutionary Schema. Available at: SAO/NASA Astrophysics Data System ArXiv. (arXiv:1712.085522 [stat.ME]), accessed May 2018, from Cornell University Library, Available at: http://adsabs.harvard.edu/abs/2017arXiv171208522L (accessed May 2018).

Lundström, S. and S. Särndal. 2005. *Estimation in Surveys with Nonresponse*. Chichester, United Kingdom: John Wiley & Sons, Ltd.

Rancourt, É., H. Lee, and C.-E. Särndal. 1994. "Bias Corrections for Survey Estimates from Data with Ratio Imputed Values for Confounded Responses." *Survey Methodology* 20(2): 137–147. Available at: http://www.statcan.gc.ca/pub/12-001-x/1994002/article/14423-eng.pdf (accessed May 2018).

Rao, J.N.K. 2011. "Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal." *Statistical Science* 26(2): 240–256. Doi: http://dx.doi.org/10.1214/10-STS346. Available at: http://www.jstor.org/stable/23059987 (accessed May 2018).

Reid, G., F. Zabala, and A. Holmberg. 2017. "Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ." *Journal of Official Statistics* 33(2): 477–511. Doi: http://dx.doi.org/10.1515/JOS-2017-0023 (accessed May 2018).

Särndal, C.E. 2007. "The Calibration Approach in Survey Theory and Practice." *Survey Methodology* 33(2): 99–119. Available at: http://www5.statcan.gc.ca/olc-cel/olc.action?objId=12-001-X200700210488&objType=47&lang=en&limit=0 (accessed May 2018).

Särndal, C-E., B. Swensson, and J.H. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Thygesen, L. and M. Grosen-Mielsen. 2013. "How to Fulfil User Needs – from Industrial Production of Statistics to Production of Knowledge." *Statistical Journal of the IAOS* 29: 301–313. Doi: http://dx.doi.org/10.3233/SJI-130784 Available at: https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji00784 (accessed May 2018).

Valliant, R., A.H. Dorfman, and R.M. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons.

Wallgren, A. and B. Wallgren. 2014. *Register-based Statistics: Statistical Methods for Administrative Data* (2nd edition). Chichester, West Sussex, England: John Wiley & Sons, Ltd.

Winkler, W.E. 2009. "Chapter 14: Record Linkage." In *Sample Surveys: Design, Methods and Applications*, edited by D. Pfeffermann and C.R. Rao, Vol. 29A, 351–380. Oxford, United Kingdom: Elsevier B.V.

Wu, C. and R.R. Sitter. 2001. "A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data." *Journal of the American Statistical Association* 96(453): 185–193. Doi: http://dx.doi.org/10.1198/016214501750333054 (accessed May 2018).

Zhang, L.-C. 2012. "Topics of Statistical Theory for Register-based Statistics and Data Integration." *Statistica Neerlandica* 66(1): 41–63. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x (accessed May 2018).

Received August 2017
Revised May 2018
Accepted May 2018