Chapter 9 — Other Key Considerations when Establishing a Statistical Business Register

Contents

Chapter 9 Other Key Considerations when Establishing a Statistical Business Register

Introduction

This chapter advises on the planning, organizational, legal and technical (IT systems) factors that will position an SBR for success.

It is intended to offer practical suggestions without being overly prescriptive. The environmental circumstances and factors within which countries will build their SBRs can clearly be vastly different. The legal frameworks for acquiring data, as well as the access to human, financial and technical resources will ultimately determine how countries can proceed.

There are however some key themes that recur, including:

- the need to build effective partnerships with data suppliers, funding providers and the users of the SBR by first making sure that the critical role of the SBR in delivering a coherent and reliable national economic statistics program is well recognized, and second by putting in place governance structures and partner engagement mechanisms that are robust.

- the need to manage the implementation and operation of the SBR in a manner that allows it to focus on and achieve its mission-critical purpose, which is to identify the population of businesses of a country so that they can be surveyed to acquire useful economic data. There are other secondary yet highly desirable roles (as described in Chapter 2) that an SBR can fulfil, such as acting as a data collection management and tracking tool. The original design and implementation plan should allow for adding in these components, but only once the SBR has fully matured in its role as a quality statistical frame.

- Generally speaking, the approach should be to maintain simplicity to the extent possible. Conceptual and technical complexities should be added in only when they serve a practical purpose, and they should never detract the SBR from meeting its larger goals.

Planning considerations:

Initial vs longer-term scoping: the importance of a modular approach

The primary function of a SBR is as a central frame for economic surveys across programs. This enables conceptual coherence, and creates the basis for an integrated economic statistics program. It is also the foundational purpose of the SBR that should be the focus at the outset of development.

The longer-term vision, however, should from the beginning also allow for the adding in of other features and components that will further enhance the SBR's value-added. The secondary outputs to be potentially developed after the SBR has become operational as a survey frame are:

- A source of register-based statistics. This requires seamless integration of administrative data, heavy quality assurance, raw data treatment and programming resources;
- A module to track respondents and response burden; (like the survey support tool described in section 2.2.4 of the manual). This module requires human IT resources to develop and maintain server and database resources. This module may or may not be efficient or necessary when using the SBR in the early stages, since population coverage remains limited and resources are directed towards the economic entities that have the largest impact;
- A receptacle for tracking survey collection outcomes and response rates, etc. (further described in section 2.2.4 of the manual).
- A 'survey feedback' mechanism that facilitates the update of frame information based on information , which is the information pertaining to frame-based characteristics such as industry classifications

The IT professionals who will architect the data structures and larger system design will in particular benefit from having the longer-term vision clearly pre-defined, thereby facilitating the addition of these modules as the SBR evolves.

Key considerations for establishing a survey frame

To re-iterate, the SBR must first and foremost be a solidly reliable listing of businesses from which statistical surveys can accurately measure the economic trends of a country. Creating it will be a challenging task, as will keeping it up to date once it is in use. The challenges entailed by the creation and on-going maintenance will be greatly facilitated by adhering to the following principles:

Do not over-extend resources in the early stages by trying to cover all types of business

While a highly developed SBR may cover a vast amount of the economic population, a new SBR must focus on covering the population that is both most important and that can be most reliably captured and reflected. The need to maximize limited human and technological resources, and to use initial funding efficiently, should limit the scope of the initial SBR population.

Reflecting the informal economy, which is highly diversified and for which no administrative data exist, cannot be a focus of the SBR development project.

Typically, for the macroeconomic indicators that are the driving objective for the statistical programs to be served, acceptable margins of statistical error can be obtained by excluding the numerous businesses that are at the smallest end of the size spectrum. Including the 'micro' businesses would add large volumes of records to be maintained, while adding only very small increments to the figures (such as GDP) being produced.

This is not to say that having these businesses will not be useful, as they can inform important policy analysis pertaining to business formation strategies, small-business financing and other micro-economic issues. This again is a desirable feature that is worth adding, but only once the core objective of adequately supporting the key indicators produce by the national accounts has been met.

Statistics Canada has gradually expanded its SBR population coverage, just having added over the past 5 years those businesses that are essentially constituted by self-employed individuals who solely pay income tax on their revenues as persons rather than as corporations.

In determining the population coverage, the conditions on the ground should be considered, as well as data that can be reliably delivered and processed from administrative sources and accessibility to administrative data. Adequate resources must also be available to efficiently obtain, process, and load administrative records.

Plan for a system that provides both live and snapshot versions of the register

It will likely be necessary to have two instances of the SBR: 1) a "live" version that can be used to receive updates so that the updating of records entailed by on-going frame maintenance activities can be instantly recorded; 2) a "snapshot version" – or "Generic Survey Universe File" (GSUF) -- to be produced from the live instance on a monthly, quarterly or annual basis that survey programs can use to create their particular sample files. This would contain all the salient fields that are stored in the SBR for a particular business record, (e.g., the unique identifier of an 'economic unit', along with its size variables, geographic information, legal information and economic activity.) Survey programs could use it to identify and stratify their sub-population of interest and draw samples. An important development task is to determine the fields that will be represented on the GSUF, as well as the fields that can be reliably populated with data from either administrative or profiled sources.

The GSUF also provides a basis for period-to-period comparisons of frame quality. "Point-in-time" estimates can be calculated and compared with one another to examine the number of records being birthed, deathed, re-classified etc. This greatly facilitates the identification of anomalies and problem records.

9.1 Governance and organizational considerations

Organizational aspects of an SBR, such as relationships with the data providers (e.g., administrative registers and survey statisticians), register co-operation (for instance with the Central Bank or other administrative institutions) and similar issues, are discussed in this section.

At higher level

In Canada, the SBR is a critical component of the economic statistics programs of the national statistical agency, Statistics Canada. The SBR is therefore held, managed and maintained within the agency. Statistics Canada is within the Minister of Industry's portfolio, and also has close ties to the Department of Finance, other federal departments, and provincial, territorial and local government organizations, to ensure that the economic statistics are relevant. Strategic direction of the program's initiatives is governed by various committees and working groups across the different levels of government. The program's governance model provides clear direction, allows for periodic review of interim results, and enables identification and execution of adjustments, to achieve the expected outcomes. The governance framework also enables transparent, effective and efficient decision-making, and supports accountability and continuous improvement of the program.

At lower lever
(Assuming the SBR resides in economic departments/statistical agencies)

Governance and organizational structure of an SBR within the economic and statistical system is important — not only for developing the program, but more so for its ongoing maintenance and support for users. The SBR should be, where possible, an independent entity with a dedicated manager.

The manager's unit should assume the following responsibilities:

- define and document all concepts, in line with international, national, and local statistical bureau standards
- plan and direct the development of SBR system processes and functionalities
- plan and implement a quality assurance program for the SBR with the goal of
    - assessing the quality and ensuring the frame's continued integrity
    - defining and producing quality measures for the SBR
    - identifying system improvements, or recommending adjustments to the training program or procedures, if required
- profile businesses to delineate those that are larger and more complex, to properly represent their production output
- ensure that businesses are classified within the proper standard industry classification
- assign or derive statistical indicators, or create statistical units, from the administrative base register to create a complete and unduplicated SBR aligned with the needs of the System of National Accounts
- validate new development strategies, specifications and procedures
- develop and deliver the courses and material to educate these SBR users: profilers, frame specialists, analysts, coders, survey divisions and collection areas (training could follow a certification process so that those wishing to access the register must first complete the appropriate level of instruction)
- support those who use the SBR data, which includes evaluating their needs for their surveys or analysis
- provide direction and support on legal aspects related to the frame such as access and dissemination
- maintain a dedicated group tasked with producing files for users and processing all files related to updating the frame.

Relationships with users

To understand the needs of its users, the SBR should have a consultation mechanism. These are typically survey areas. The SBR team will need to meet teams from survey areas regularly to understand the changes and requirements of the economic world. Not every request need be accepted: the SBR serves as one frame for all its users, and balancing available resources with requests is challenging. The manager of the SBR must always be aware of the role of the frame in the context of the larger statistical program.

9.1.1 Legislative framework

As stated earlier, access to administrative records such as corporate tax returns, business registrations, payroll deduction or sales tax remittances will be fundamental to building a centralized SBR. Many countries legally require the provision of such records for the purpose of compiling official statistics, and this authority must be fully leveraged to build an SBR where it exists.

In doing so, it will be important to have detailed agreements – usually formal Memoranda of Understanding (MOUs) -- in place with the administrative agencies in order to clearly spell out the terms for the sharing of administrative data, stipulating the types of data required and the likely treatment of the raw data. The agreements should also specify data security measures and the means of transmission to be used.

These MOUs are important because they establish a framework of general rules and procedures for interdepartmental data exchanges, and they specify data protection measures.

9.1.2 Funding

The funding model to be put in place for the SBR will itself be an important aspect of the governance and decision-making process, and should be given careful consideration at the outset.

It is unlikely that the costs to create and maintain the SBR, which is a kind 'public-good' resource to be used by a variety of client programs for a diversity of purposes, could be recovered entirely on a fee-for-service basis. There may be some individualized activities, such as preparation of special data requests for individual clients, for which the incremental costs could be recovered directly from the clients being served, but the development and maintenance of the system are generalized expenses. Therefore, the majority of the costs will likely be funded through subsidies from the public purse.

The key questions then become: 1) who should make the budget request? and 2) who should control the budgeted funds once they are received? The answers will depend on the organizational governance structures in place, but the objective should always be to ensure that the allocation of funds is such that the larger objectives of the economic statistics program are optimally fulfilled. It will therefore be preferable for the budget to be controlled by those who govern the larger statistics program. The manager of the SBR itself can then make specific budget decisions under the direction of this larger governance structure.

Developing the SBR consists of three distinct stages. The last is maintenance (or post-build operational requirements): these should be accounted for during the development and funding plans. We will briefly review each stage as well as its associated tasks and funding components.

Conceptual development stage (pre-build)

This stage requires initial funding and resources for the following pre-build tasks:

- defining the initial scope of a SBR development project
- defining the initial population coverage
- analyzing and determining all of the data input needed to populate the SBR to create a coherent, accurate and timely frame
- analyzing and determining the necessary stratification variables of the SBR
- defining the most viable means of receiving business inactivation and cessation information, to be carried out during the development phase of the SBR development project
- determining the availability of raw data and drafting MOUs: the cost of this phase is greatly influenced by the accessibility and usability of the data input
- drafting MOUs and negotiating legal, ongoing data-sharing agreements with other departments
- outlining the processes that will validate and treat the data before it is loaded into the SBR
- defining components for future developments, using the modular approach explained in the introduction of this chapter.

Development phase
Development funding will depend on the decisions made during the pre-build phase. A modular development approach is advantageous because development can be phased in and compartmentalized module by module. For a fully operational SBR, the following modules should be funded and fully functional for the targeted coverage of the initial SBR:

- database
- batch load/update process and online maintenance
- output (survey interface)
- quality assurance

Investment plan or operational/maintenance phase
The SBR is a continuously evolving entity. A long-term vision, supported by a senior management team, is needed. In a statistical agency, a senior steering committee must oversee the needs and usage of the SBR. This will ensure that it evolves with the agency's data requirements. A 10-year investment plan should be outlined to prepare the funding schedule. This plan should be reviewed annually and adjusted to new business requirements.

9.1.3 Human resources

What resources are needed will depend on the financial support that each country dedicates to developing and maintaining an SBR with a quality level that supports the economic program. Investing in a high-quality SBR that is timely, accurate, coherent and accessible (i.e., user-friendly) will result in lower costs later in the economic program, by allowing for a simpler sampling system, lower collection costs, improved response rates, higher quality estimates and lower response burden.

Ongoing operational requirements and organization
The operational structure of a completed SBR, and its attendant funding structure, should be clearly set out in advance to ensure that the build project will not lead to an ineffective or underutilized SBR. The operational requirement for a viable SBR, which is at the heart of an integrated statistics program, is explained later in this document. A 'sectional' structuring has been devised to clearly delineate tasks, responsibilities and funding commitments. The SBR operational team can be divided into three groups with distinct roles (though opportunities for efficiency and collaboration can be pursued): IT and systems, quality assurance and data coherence, and operations and data maintenance.

IT resources
Maintaining the SBR will require dedicated IT staff. As further detailed in Section 9.2, at least one database administrator (DBA) will be required. The DBA will interact with the database management system processes and tables that make up the register, and ensure that the tables are accessible and available for the GSUF outputs. The DBA will be responsible for ensuring that the database, and the server to a lesser extent, is functional. The DBA will also need to ensure that the database can be updated and queried as required.

An IT team for the SBR is also required, to ensure that the systems and software needed for extracting data (i.e., the GSUF and specific sample files) are properly programmed and optimized. This team will also maintain, and possibly advance or further develop, the graphical software that will enable SBR staff to easily access and update the content of the SBR. This may initially be restricted to packaged server software (e.g., Microsoft SQL Server Management Studio). The SBR IT staff will also ensure that certain derivation jobs (for codes such as SNA) are programmed and run correctly.

Given the limited scope and functionality of the initial SBR design, one or two dedicated IT personnel should be sufficient for these tasks. Funding will also need to be secured, to ensure that software licenses are maintained and that equipment — servers, PCs, networking equipment and the networks themselves — can be maintained and replaced when needed.

Sufficient numbers of highly qualified people with knowledge on the business processes (i.e., the data model and process model of the register) are essential. They must also be competent with specific IT requirements, to develop and maintain the register application, and to test newly developed register features.

IT specialists should include

- database/data model specialists
- programming language like SAS and SQL specialist, to process large quantities of data
- .NET specialist for the SBR interface.

Quality assurance and data coherence staff

As the SBR is being established, and once established, the human resources that perform the raw data analysis must be extended, to further assess the Quality Assurance (QA) requirements and develop a strategy for QA controls.

A person or a team of frame specialists, ideally with profiling experience and strong technical skills to perform data and data processes validation, analyze frame data and investigate potential quality issues. A QA and data coherence team will be maintained as part of the SBR's ongoing operation.

This team should comprise

- data source specialists to create specifications or use cases for the development of batch processes
- specialists in business register concepts and business structures
- business survey interface specialists.

This team will be responsible for

- providing specifications to the IT team on the variables to be populated on the database tables, as well as maintaining the content of the tables and submitting new or modified specifications as needs arise and shift
- receiving and reviewing external administrative data from other departments
- conducting internal analysis and coherence checks on the frame information
- reviewing requests from methodology, subject matter and other SBR users (including the profilers) for improvements or modifications to the SBR metadata, variables and outputs, and prioritizing these requests, as well as creating specifications for future development and improvements
- liaising with subject matter and methodology clients to ensure that they understand SBR processes and outputs, and that they are submitting and receiving files from the SBR as needed
- ensuring the quality of a population frame before it is released to users
- identifying significant errors
- ensuring timeliness, coherence, accuracy of updates
- documenting concepts, processes or coverage changes over time.

Operations and data maintenance section: maintenance of SBR information
A SBR operations and data maintenance section is needed, whose main task is performing profiling activities on the frame to expand on and fill in the information received from administrative sources.

The SBR maintenance process through profiling can be organized so that profilers are assigned a certain number of large, critical enterprises that they are responsible for updating and reviewing. They receive signals from various sources on cases that may require updating, and periodically perform proactive profiling to augment the reactive profiling. The operations team will investigate and correct incoherencies on the frame that are identified by the QA team. Since the largest enterprises can be grow and change dynamically, sufficient resources are needed to continually maintain the accuracy and relevance of the frame information.

A SBR that remains static after its creation is not useful. A realistic, well-funded maintenance strategy is a crucial aspect of the SBR development project. The exact nature of the signals and processes that will prompt a profiler to review a case can be examined in more detail as development progresses. When a SBR is new, the operations section might determine a list of the most critical enterprises — based on public data, trade publications and information from other departments — and set a schedule of proactive profiling for these cases. Securing and committing full-time personnel to these tasks is essential to ensure that the frame remains relevant and accurate. The exact size of the team will depend on the coverage of the frame at its outset.

9.1.4 Relationships with other registers

The SBR should be established knowing that it may be linked to other registers or lists. Thus, from the start standard definitions and concepts must be used, and all the most commonly used unique identifiers for each type of businesses must be included.

> For example, in Canada every business that registers with the Revenue Canada Agency is assigned a Single Business Registration Number (SBRN, commonly known in Canada as the BN, or Business Number). It was then essential for the SBR in Canada to have a clear understanding on how to relate/link the BN with the Statistical Identifier(s) of a business and to include the BN as part of the SBR as a key identifier

Developing a flexible SBR system that enables linking or even permanently storing the various concordances of identifiers may prove to be very useful. As well, satellite systems may need to maintain their own identifiers.

9.2 IT Solutions

The scope of the coverage, as well as the number of variables (real, estimated and derived) should have been determined during the pre-build phase. Thus, choosing an IT infrastructure and developing the requisite processes to run on it should flow more easily during the build phase. This section offers recommendations and notes on the likely IT infrastructure and programming requirements for the build phase. Cost and resource requirements can begin to be inferred from these notes.

IT Development
The IT development should be considered in two parts:

- acquiring an initial database infrastructure, keeping in mind the modular approach to growing the SBR — the infrastructure should be expandable and flexible

- the programming and process development that will be required in order to create
    - the tables that will make up the database
    - the main statistical outputs of the frame (GSUF, sample files, etc.)
    - the programs and tools required to query and update the database.

When establishing an SBR, many technologies can be used. The solution used should take into account scalability, cost and maintenance. As the SBR will continuously evolve, the choice of technology should be flexible enough to evolve with new requirements.

9.2.1 Database

A relational database should be considered when establishing an SBR. This will enable properly segmenting the areas that make up the whole, which will in turn enable a proper maintenance strategy. The database should be scalable: there may be a very large number of observations to process. SQL and Oracle are two of the platforms that could be used to establish a SBR. These platforms also enable proper authentication of users.

Collection modules and frame modules

The SBR, in its capacity as an economic frame for sampling, must be seen as a database that stores both the frame information itself (e.g., unique identifiers and stratification variables) as well as the 'collection' information. The collection information that Statistics Canada's Business Register Division (BRD) stores on its frame includes a record for all units sampled, information on where those units were sent for collection and the outcome of the collection effort. This dovetails with the response burden module that BRD maintains. The response burden module provides an agency-wide view — across economic statistical programs — of the reporting burden that the statistical agency places on enterprises, and a record of the efforts to mitigate this burden (e.g., tax replacement).

A response burden module might be developed at a later stage of the SBR development project, but, to start, a simple collection module should be developed and maintained. A key benefit of using a relational database as the foundation of the SBR is that tables can be created, developed and maintained apart from one another. This provides great flexibility regarding resource requirements and allocation, the sophistication of the various tables, and the ability to build on different modules at different rates. While BRD recommends that initial resources be devoted to developing the actual frame and its main outputs, simple collection tables can be devised during this phase to track which units on the SBR will be sampled most heavily. This will also allow for future resources to be allocated more efficiently as usage is tracked.

DBMS options

Relational Databases and SQL

The modular approach to building the SBR could be aided by using a relational database structure that would enable creation of different tables and modules in isolation from one another. Using a mainstream SQL database solution such as Microsoft SQL Server offers other benefits, including predefined software server applications (e.g., SQL Server Management Studio) and international standards for querying the database (SQL languages). IT infrastructure costs can also be rigidly defined once the needs for initial coverage and comprehensiveness are determined. This relationship, between coverage/complexity of the frame and the infrastructure required to maintain it, can be viewed as proportional — coverage and complexity can be adjusted in line with the resources available for server and maintenance budgets. A relational database model provides a high degree of scalability which is precisely what this modular model of SBR development calls for.

Storage Metrics and Server Costs

The cost of IT infrastructure will also be affected by the database storage metrics that become necessary vis-à-vis the coverage levels. Metrics will need to be determined for access speed, packaging density, power efficiency and the level to which data can be compressed. The cost of the infrastructure will rise in line with these metrics. Other factors that can be controlled for include the number of users that will need to be able to access the database at one time, how many people will have write access to the database, the number of tables the database will house, and the complexity of data constraints (i.e., the number of parameters around what values are accepted for various variables). Data constraints are vital: they will be closely linked to automated quality control. A DBA resource will be needed to maintain the servers and IT resources and to help maintain their data tables.

It is recommended that at least two separate databases be maintained, one for the actual frame and one for testing, development and troubleshooting. Two servers is ideal, to maintain mirrored copies of their frame for security and stability.

> In Canada, the SBR is now maintained in about 30 databases on five separate servers. For example, the BRD keeps its frame on a separate database and server from the snapshot of that frame to which users have access; this helps limit data corruption. Initial requirements should not be as complex as BRD's, and using SQL Server databases enables addition of servers and databases as resources and growth dictate. Another benefit of a relational database is that users can link up tables across several BRD databases to perform analyses and track data.

Simple Database solutions

There are other, lower cost options for simple database solutions, but it is BRD's view that the cost benefits of these solutions do not mitigate the loss of functionality and flexibility (for usage and future growth) that a relational database offers. For example, MS Access could be used to run a limited frame, but with several shortcomings:

- the inability to program sophisticated data constraints
- limitations for users in how they could query and update tables simultaneously
- constrained ability to proceed with a modular approach that would enable gradual introduction of databases and tables
- limited security controls.

A relational database setup that can be grown over time is recommended.

Quality Assurance and Coherence Processes

A great advantage of a relational database system is that data constraints can be placed on the tables. This ensures that inputs (either from profilers or from administrative data) must meet certain criteria in order to be loaded. This will allow for automated QA checks on data when they are loaded. Furthermore, the division of different types of information into multiple tables allows for data segregation: thus, analysis can be atomized further. The IT development team should be contracted to develop certain data constraints. As well, they should design coherence programs that can be used to determine outlier data in tables and validate the internal coherence of certain information (for example, that the establishment revenue in a structure is not greater than the total enterprise revenue). Constraints on data should also be implemented as a control on the quality of the data being received and loaded from external sources.

Programming requirements

Database Construction and Organization
To establish the SBR, an IT development team is needed to design the tables and define the constraints of their frame. At least two database environments must be developed at the outset of the project. Designing the database tables will involve developing an architecture that defines which tables will house which variables and the manner in which those tables should interact. The IT team must also determine how data will be backed up and secured, and must create a set of database roles to be assigned to users as needed (e.g., who will have write access to the frame.)

Another important task in table design is creating unique identifiers for the records on the SBR, both to identify specific records and to identify the hierarchical relationships between the records. The latter is key: it enables the GSUF to denote the parent–subsidiary relationships of enterprises, as well as the operational links between enterprises and establishments.

Graphical interfaces and database server software
The manner in which users will interact with the SBR is a crucial decision in the build process. Two things should be considered: the types of users and their requirements, and the relationship between the Frame itself and its outputs. Users of the SBR will include

- employees of the SBR maintenance and profiling team, whose task it is to profile enterprises and ensure their information is accurate and up to date
- subject-matter users, who want to look at the enterprises and structures on the SBR that impact their own statistical programs to review, and possibly update, their universe
- methodologists and statistical staff who will analyse the frame and population to select samples.

Each group of users has different requirements. The methodology group will be most interested in analysing and parsing the outputs from the frame (the GSUF) to create their samples. Profilers and subject-matter users, however, would greatly benefit from a graphical user interface (GUI) that would enable them easily to search, browse and update entities on the SBR. A GUI would also minimize the need for all staff to understand SQL and database querying. A GUI however, would present large costs for development and maintenance (including design, coding and additional server infrastructure). Initially, running database server software like SQL Server Management Studio (as well as SAS and SAS EG) to access and query the data on the SBR may be more economical. The SBR could be made operational and start producing the outputs required more quickly and cheaply than if a GUI were to be developed. Without a GUI, there is still plenty of capacity to group and analyse enterprises based on the database tables and the GSUF output: thus, the cost-benefit at the outset of SBR development is tilted against a GUI. Keep in mind that the GSUF is, ultimately, the main output of the SBR: the GUI is a tool that will eventually enable statistical staff to more easily browse, profile and update the entities on the frame. The GUI is not a necessary SBR function.[1]

Statistical extraction processes (GSUFs, sample files)
Methodology and programming will need to be developed to produce the extractions that make up the main SBR outputs. The extraction process will require relatively little programming, as it is a derivative product of the database tables themselves. The main effort

1. A GUI does, however, offer benefits in ease of use and data quality control: using SQL and allowing users to write data directly to the frame could create the potential for more errors than using a GUI with field edits built into the system.

will be spent determining which variables will be available on the GSUF and which will be available on the sample files. The exact design and types of tables, and the unique identifiers to be used, should have already been determined before the design of the extraction processes. A main point for design of the sample files extraction process will be to determine what information subject-matter/methodology must provide as an input (e.g., unique SBR identifier, type of survey, reference year) to yield the sample file output from the SBR with their desired variables. Please refer to the GSUF and sample file examples included from BRD.

Creating a data repository to archive the GSUFs and sample files that are created is recommended, as these are the primary outputs of a SBR. Maintaining these files is important from a data management perspective. They also offer a key input for future analysis and development work.

9.2.2 Environment IT solutions

IT solutions should be established to properly manage the deployment of new features on the system. The suggested IT solution proposed below is composed of five distinct environments.

- The production environment should be a dedicated version of the system, including the active data that are being updated.
- The practice environment is the same code as production, but the data are not continuously updated. This environment is typically used to perform updates that mimic what would happen in production. It enables users to try and see how their changes would impact the overall process.
- The user acceptance environment enables testing of new programming functionality before moving the code into the production and practice environments. It typically has fictitious data that are developed solely to test various scenarios.
- The development environment is dedicated to systems programmers, who use it to test their own programming. Once code is system-tested, it can be moved into the user acceptance environment to be tested by users.

It is important to consider how all other systems link to the SBR when establishing this IT solution/structure, as it may be important to link these other system to their corresponding SBR environments. For example, end-to-end testing across systems may be required: for this, the SBR in the user acceptance environment would need to be linked to the user acceptance version of the ‘linked system’.

- The analysis environment is dedicated to analysts performing quality evaluation and simulation testing. This environment offers analysts access to a mirror image of the data available in the production environment without disturbing the production processes.

9.2.3 Programming Language

A programming language should be chosen that is common enough that knowledgeable staff and training are available. As well, the language should have a demonstrated a long shelf life. This decision is crucial, and should be researched thoroughly. A poor choice could lead to premature redesign and costly maintenance processes.

9.2.4 Tools/software for data matching

The results often touted for automated tools or software for data matching tend to be overly optimistic. Our experience has shown that, without a clear unique correspondence between two sets of unique identifiers from two or more sources showing a hard link between those

data, probabilistic matches often yield erroneous matches when the matching rules are ‘too loose’ and a low percentage of matches when the rules are ‘too rigid’.

9.2.5 Job-scheduling software

Job-scheduling software is often overlooked. When establishing an SBR, understand that maintenance, updates and extraction of data will be required beyond the manual process that a GUI would provide. The administrative data that will be used to update the SBR usually requires a great deal of time to process. To speed processing, it is usually done when no one is accessing the base — typically at night. For efficiency, job-scheduling software will enable proper management of the nightly updating process. The software should be able to monitor processing and remotely notify technicians of any problems to be resolved before users access the system the next morning.

9.2.6 Documentation

Documentation related to programming the system is needed to ensure the long-term function of the SBR. Your IT team should invest significant time properly documenting each module and process, both during initial development and as the SBR evolves. This will help current staff understand the changes they need to make: these documents should be detailed enough for a new programmer to continue where the previous one left off. Wikis enable free-form information entry over time.

9.3 Data retention

As part of establishing the SBR, a strategy on data retention must be spelled out to understand future cost and space requirements while accounting for user needs and legislative restrictions. The strategy on data retention will depend on operational and analytical needs. Frequency and details of the archiving process should be determined and documented in a directive. A privacy impact assessment might be required.

Tracking changes first

As we determine the data retention strategy, we must also define how we will keep track of the changes made to the SBR data. The SBR is constantly being updated, which raises the question of to what extent, and how, historical information should be stored. One approach is to always add data, never replace it, record the date and time each item of data is created and, if it is subsequently updated, the date and time of the update. (with or without the effective date corresponding to the date the data changed in reality. This approach enables the creation of a view of the SBR as of any past date and time. The alternative, much simpler approach is to take periodic snapshots of the database and keep these for as long as seems necessary. To satisfy all operational and analytical needs, it is suggested to consider using both approaches.

9.3.1 Operational

Operational data retention should be frequent enough to offer users a reasonably comprehensive view of the past. A monthly snapshot of the main data going back two years would be a starting point: it would offer users sufficiently frequent data for their immediate needs.

Keeping track of frame updates is also recommended. An automated log creation should include the update date and the source of the information used to update (i.e., the name of the administrative source). This information may prove useful for troubleshooting issues with a specific maintenance or update process, and may serve as the initial source explanation or documentation of changes for SBR users.

9.3.2 Analysis

Data retention for analytical purposes addresses the needs of statistical users. An annual picture of the main data holding should suffice for most users. Some countries retain these file indefinitely, as they are a great source of information for longitudinal studies.

In Canada, for example, a complete copy of the 'production' SBR database — called a 'snapshot' —is taken just prior to the first day of every month. A GSUF, containing every statistical entity, is created from these snapshots every month. Although GSUFs are primarily used for sampling, normally soon after the GSUFs are created, the GSUF is retained for an extended period for analysis purposes. The table below shows the current retention period for each GSUF in Canada.

Monthly GSUFs	**Retention period**
January	Indefinite
February to December	24 month

Data are important, but understanding them is even more important. As the SBR undergoes constant evolution in concepts, coverage and methods, the data retention strategy must also contain documentation explaining these changes. Each file stored should have a full account of major changes attached to it. This will enable future analysts to understand the data. The quality assurance team may also use the log information to assess the quality of certain maintenance processes or sources of SBR updates.