INTERNATIONAL SEMINAR ON
PRIORITY CHALLENGES
IN
PENSION ADMINISTRATION

TOKYO, JAPAN

JANUARY 20-22, 2010

Ensuring
Appropriate Information Management
& Data Quality

Mr. Jerry Berson
Assistant Deputy Commissioner, Systems
United States Social Security Administration

Ensuring Appropriate Information Management and Data Quality at The United States Social Security Administration

Introduction

The Social Security program of the United States (U.S.) was established during the Great Depression of the 1930's. In June 1934, President Franklin D. Roosevelt created the Committee on Economic Security to study the situation and make recommendations on how to improve economic conditions in the U.S. The committee report, completed in January 1935, was submitted to Congress by President Roosevelt resulting in the passage of the Social Security Act of 1935.

Signed into law on August 14, 1935, the act contained a number of provisions, including the creation of an old-age insurance program that would provide income to retired workers age 65 or older. The act also created the Social Security Board (SSB), later renamed the Social Security Administration (SSA), to oversee the program that was to be funded by withholding a small percentage of employees' pay, supplemented by employer matching.

From the outset, the Social Security program presented daunting logistical and operational challenges. Set to begin in January 1937, there was much preparatory work to be done and no organization in place to do it. One of the immediate questions was how to collect and track the contributions of each covered employee. To aid in the record keeping process, the SSB decided to create a unique identifier, called the Social Security Number (SSN), to be issued to all workers covered by the Social Security program. The issuance of SSNs and their subsequent use to identify workers and track their contributions would become the beginning of today's Social Security information management system.

Because the newly created SSB lacked the means to undertake the huge task of registering workers, the U. S. Post Office was enlisted, and beginning in November 1936, the Post Office began the process of distributing millions of applications to workers across the nation. The completed applications were then returned to the Post Office, where postal workers typed the SSN cards and mailed them to the applicants. Over 35 million Social Security cards were issued through this procedure in 1936-37.

The registration information collected by the Post Office was forwarded to the SSB's new processing center in Baltimore, Maryland. The SSB made Social Security's first "information management" decision when they chose to base the program's record keeping operations on the most advanced information technology available the time - punch card based Electronic Accounting Machines.

The Social Security program has been repeatedly expanded over the years to broaden the coverage and add new categories of benefits. In 1939, benefits were added for the spouse and minor children of a retired worker (called dependent's benefits) and survivor's benefits were added to protect a family in the event of the premature death of the worker.

The Social Security Amendments of 1954 and 1956 established a disability insurance program to broaden the protection against economic insecurity. In the1960's, the Medicare program extended health coverage to Social Security beneficiaries aged 65 or older and in 1972, to those receiving disability benefits.

In 1974, Social Security assumed responsibility for administering the new Supplemental Security Income (SSI) program. This needs-based program for elderly, blind and disabled individuals replaced and standardized a series of state administered programs that had been created by the original 1935 Social Security Act.

In 2004, a prescription drug program was added to Medicare and Social Security was given responsibility for administering the portion of the program dealing with subsidies for low-income enrollees.

Throughout this period, the SSA strove to sustain a high level of technical achievement. The punch card based systems of the 1930's and 40's were phased out in the 1950's when SSA became one of the first organizations to enter the computer age, transferring data from punch cards to magnetic tape and developing batch computer systems to run on "mainframe" computers with 20 kilobytes of memory.

This tradition of information management innovation at SSA has continued through multiple generations of technology, always driven by an enduring commitment to provide the best service to the American public in the most efficient manner. This spirit is evident today in the agency plans to fully leverage the capabilities of the Internet, adopt new service-oriented technologies, and deploy a robust and resilient technical infrastructure hosted by geographically dispersed data centers.

SSA's data and databases form the foundation for the agency's massive business processes that provide more than $600 billion a year in benefits. SSA's data, separate and apart from its original purpose, has become a national asset as the Social Security Number (SSN) has become important far beyond its original purpose. Individual citizens, States, Federal agencies, loan officers and human resource departments across the nation all rely on SSA's data to verify citizenship status, eligibility for employment and a wide range of services, from driver's licenses to disability and retirement benefits. U.S. citizens should rightly be confident in the accuracy and integrity of this data, a responsibility that belongs to SSA alone.

SSA is unique in other ways as well. All large government agencies have important data demands, and some agencies deal with data volumes that exceed that of SSA. Few, however, rely on the accumulation of data collected over the past 72 years to perform their day-to-day core business functions. In 2009, processing Social Security claims for retirement, survivors or disability benefits is based, in part, on information captured when a claimant's SSN was issued and on the claimant's accumulated lifetime earnings, all stored in the agency's databases. In fact, some of the SSNs issued by the Post Office back in 1936 and 1937 are still in active use, serving as the identifier for Social Security beneficiaries receiving monthly payments from SSA.

Data quality issues, such as missing or invalid data are more prevalent for older data. As one would imagine, "legacy data" captured years ago was not subject to modern data validation techniques and depending on when data was captured, over time, it may have been stored on numerous generations of technology, each with its own storage formats and coding schemes. The data conversion processes between generations of technology provided additional opportunities for the introduction of data anomalies, thus adding to the complexity of today's processing challenges. The continued relevance of "historic" data distinguishes SSA from other agencies and emphasizes the importance of the SSA's role as the steward of this invaluable information. Fortunately, the SSA has long focused on preserving and protecting this data. Today, as SSA's mission is rapidly evolving and expanding to include many new interactions, it is essential that the agency's data strategy be proactive and forward leaning – continuing to do well what SSA has always done well, but seeking to excel in new ways. These new ways increasingly require the sharing of data throughout SSA and with external partners without compromising security or quality.

Major Programs & Workloads

Over the years, SSA's responsibilities have expanded into a collection of major programs and supporting workloads that, working together, provide an important system of financial protection for the American public. The table below is a summary of the agency's major programs and selected support functions, the date they began, and annual workloads.

Program/ Workload	**Year Initiated**	**Annual Workload**
Enumeration	1936	5.1 million SSNs issued (more than 456 million issued to date) 10.6 million name changes and cards reissued 1.25 billion verifications of name and SSN
Earnings History	1937	250 million earnings reports totaling over $5 trillion
Retirement & Survivors Benefits	1939	3.9 million retirement applications
Disability Benefits	1956	4.5 million disability applications
Supplemental Security Income Benefits	1974	290 thousand applications
Document Management Architecture	2004	165 million documents added (more than 490 million added to date) 310 million pages added (more than 1,240 million added to date)

Current SSA Challenges

SSA plays an increasingly vital role in the economic security of the United States. Over the years the workload has grown, not only through population growth, but also through legislative expansion of the agency's mission.

As the agency has done throughout its history, it must continue to aggressively pursue appropriate information management opportunities in order to meet the challenges the agency faces going forward. It is clear that challenges from both external and internal forces will be drivers of change. Four of these areas are outlined below.

Rapidly growing workloads - The SSA workloads have been growing rapidly and are projected to continue to increase substantially in the near future. The first of the 80 million baby boomers have already applied for Social Security retirement benefits. In addition, over the past few years, the aging of the baby boomer generation has also driven disability workloads significantly higher, resulting in what is sometimes referred to as the "Disability Wave." Recently, the combined impact of the aging baby boomers and the downturn in the economy has resulted in a dramatic increase in the number of disability claims filed, creating a backlog in SSA's disability adjudication process. It is anticipated that over the next decade, workloads will jump as retirement claims increase by more than 40 percent and disability claims by an additional 10 percent. Other workloads are also rapidly expanding. In collaboration with U.S. Citizenship and Immigration Services, SSA recently completed a major upgrade to the E-Verify system used to verify work authorization based on information contained in SSA's SSN database. Since its inception as a small pilot in 1996, this system has grown to over eight million queries per year.

The retirement wave of SSA employees - Many SSA employees have made a lifetime career at the agency, creating an environment rich with experience and knowledge of the agency's key information technologies and business processes. However, many of SSA's senior technical leaders are, or will become, eligible to retire in the next ten years. When they leave they will take with them critical technical and business knowledge essential to SSA's existing systems. Accordingly, SSA has initiated succession planning and knowledge transfer processes that focus on the perseveration of critical technical and business knowledge.

Legacy technologies & aging technical infrastructure - As SSA gears up to meet the increasing demand for services, it must also continue to modernize the agency's technical environment. Although the current systems are meeting today's demands, major portions of their infrastructure were deployed in the 1980's. The infrastructure put in place was forward looking for the time and has served SSA well for more than two decades. However, portions of this infrastructure are nearing the end of their useful life, and SSA needs to embrace more modern, industry-proven technologies, architectures, frameworks, and processes.

Fortunately, SSA has already begun that process on several fronts. Today, many of SSA's business applications are written in a useful, but dated, language called Common Business Oriented Language (COBOL). SSA is transitioning from COBOL to web technologies for front-end applications. The agency is also making good progress on a multiyear data migration to an industry standard database management system to replace the agency's older homegrown data storage system known as the Master Data Access System (MADAM).

To accommodate SSA's growing computer operations, SSA has recently established a second fully functional Data Support Center. This center supports portions of the agency's critical and non-critical workloads, and the built-in redundancy it provides also enhances the agency's disaster recovery capabilities. Since the existing National Computer Center is more than 30 years old, the agency is also pursuing plans to replace it with a new facility with sufficient capacity to support increasing workloads and meet future technology and security needs.

Emerging technologies and rising customer expectations - The rapid deployment and public acceptance of new technologies have elevated client expectations regarding service speed, availability, and accessibility. The general public is very familiar with the Internet and many enjoy broadband technology which ensures that, regardless of what website they visit, they can access the information they need quickly. Users now expect to be able to access all of their information online, just as they can access their banking information online. This results in the expectation that government agencies should likewise be able to accommodate information requests faster. In response to this known expectation, the agency now offers numerous online services. For example, it is now possible to file several types of claims, request services and provide information directly to the agency over the Internet. Examples include applying for retirement benefits, estimating retirement benefits, requesting a social security statement, applying for disability benefits and several other services.

Leveraging the full capabilities of Internet-based business systems will require a robust and secure infrastructure that can scale to handle a large numbers of concurrent users. It will also need to be very reliable, with scheduled availability based on a thorough analysis of business need, workload metrics, and thoughtful evaluation of technical alternatives.

SSA's Information Management Priorities

SSA leadership is committed to both sustaining current levels of efficiency in the agency and increasing agency capabilities over the next five years. These details are documented in both the Agency Strategic Plan (ASP) and the five-year Information Technology (IT) Vision. The ASP for fiscal years 2008 – 2013 places a heavy emphasis on reducing the current disability backlog and improving the disability hearing process, as well as improving retiree and all other core services provided by SSA. The overall intent of the plan is to make aggressive changes to keep pace with the changes in demographics, while also preserving the public's trust in Social Security's programs. Recognizing the key role of IT, the ASP sets forth a strategic principle to guide the agency's IT direction: ***"SSA will use innovative technologies with a robust infrastructure to meet the changing needs of the American public."***

The agency's IT Vision focuses on information management and defines three strategic imperatives that align directly with the ASP goals:

- Changing how we do business
- Building a stronger IT foundation
- Revamping software and databases

The rapid growth of agency workloads has caused marked increases in demand on IT infrastructure. The chart below shows recent and projected future growth in key areas.

Information Technology Growth Chart

Growth Measures and Projections

IT Infrastructure Segment	2001	2009	2014 (est.)
Network Bandwidth - millions of bits per second	2,216 Mbps	13,388 Mbps	18,589 Mbps
Mainframe Storage Capacity - terabytes	12	1,054	3,300
Mainframe Processing Capacity - GP MIPS	4,400	50,182	62,000
Transaction Volume - Average per day (million)	27	70.4	130
Workstations Supported	107,000	130,448	146,500
Internet Transaction Volume - Page views per quarter (million)	101.2	1,000	3,900
Production DB2 database instances	5	492	900
Total number of DB2 rows (million)	100	46,000	200,000
Telephone System / National 800 Number calls (million)	77	86	92.1

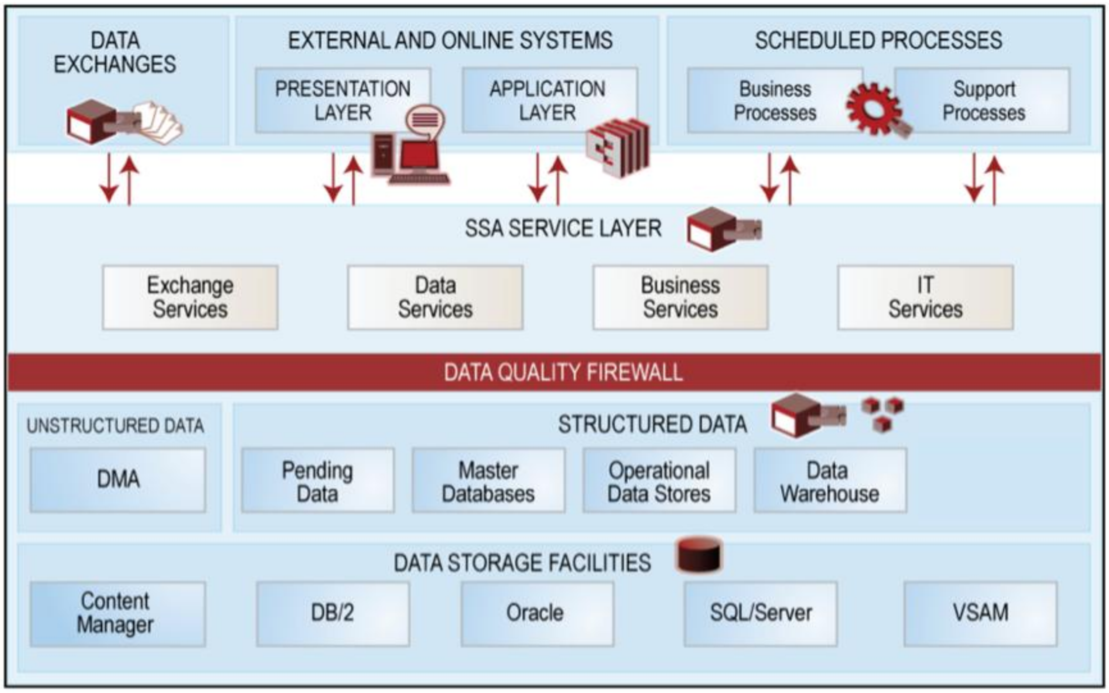
Information Management Direction for Data and Databases

SSA recently engaged industry experts to assist in developing a new Enterprise Data and Database Strategy for the agency. This strategy calls for aligning, strengthening and leveraging SSA's data related capabilities in a way that best supports Social Security's mission and the goals and objectives articulated in the ASP and IT Vision. It also calls for SSA to employ modern, industry proven data and database practices to improve the effectiveness and efficiency of the agency's business processes, thereby improving its service to the American public.

In recognition of the national importance of the agency's data holdings, the strategy provides for safeguarding the quality and integrity of SSA's data and databases. It also focuses on improving the throughput and agility of the agency's systems development processes to facilitate and expedite the delivery of new capabilities into production. New technology, new management models, and improved data delivery techniques are also described in this strategy, including the concept of data delivered as a service. The strategy will also facilitate cross-component collaboration as the agency works to put in place the envisioned future data and database environments. Taken together, implementation of the initiatives articulated in this strategy will significantly advance the agency's ability to successfully meet the challenges of its rapidly growing and increasingly complex mission.

Note: The information contained in the following sections are excerpts from the Enterprise Data and Database Strategy document and are included because of their pertinence to the topic of this paper. The strategy document has been recently completed and is not yet fully reflected in SSA's plans. The document describes future capabilities and proposes processes to achieve them. While work is currently underway on some of these capabilities, others will require funding to move forward.

The conceptual illustration below depicts a data-centric view of the proposed future architecture.



The above diagram depicts salient aspects of the future environment including both logical and physical aspects of the architecture as well as the enabling principles, support processes and procedures that will be key to its success. The data centric future state will be based on an Enterprise Service Oriented Architecture (SOA) approach. This will enable the creation of a layered architecture that aggregates similar functions such as data presentation and capture into one layer and applications processes into another.

The following objectives are enabling principles:

- Support an enterprise SSA service layer that capitalizes on SOA principles by developing data services
- Separate data presentation and capture from application processes
- Implement program-neutral data capture
- Encapsulate key business processes and expose them as business services
- Insulate applications systems from physical databases through data services
- Extend the conceptual enterprise relational model to the logical and physical levels
- Foster data standardization through enterprise data governance processes
- Ensure high quality data through data profiling, analysis and data quality firewalls
- Facilitate data exchanges through standards based (e.g., XML) interfaces

- Ensure high quality software by using sanitized production data in end-to-end testing

The creation of an SSA Service Layer will be an important feature of the new environment. It will provide an enterprise level integration interface that will support encapsulation of certain key business functions and expose them as business services to be invoked where needed. This will facilitate logical integration and reuse of business services across architecture layers and offers the potential to simplify and streamline the implementation of future business changes. If a change can be isolated to a business service invoked by multiple processes, those invoking process will not require change.

The Service Layer will also enable applications to obtain the data they need via data services. This will decouple software from the physical considerations of data access and management.

Coupled with other data quality analysis such as data inventorying and data profiling, “data quality firewalls” will conduct the necessary checks to ensure only valid data is accepted into the agency databases. This step will attack the root cause of data quality issues and prevent future instances from occurring.

The Presentation Layer will provide service to client contact employees and also directly to clients over the Internet. It will provide both the ability to view and enter data through web-based applications. Finally, the Data and Database Layer is the foundation for the entire data architecture. For managing unstructured data, such as images, video, etc., SSA plans to expand the use of its enterprise facility, the SSA Document Management Architecture (DMA), a robust platform based on current state-of-the-industry technology. In line with SSA’s plans to move away from its legacy data management tools and focus, instead, on industry standard methodologies, SSA’s enterprise structured data holdings will be housed in modern, robust, commercial Database Management Systems (DBMS).

There will be numerous benefits from this architecture. The stratification of this architecture will facilitate standardization. For example, the separation of the data presentation and capture processes from application processes will enable data to be captured once, based on agency standard definitions, as defined in the Enterprise Data Model, rather than multiple times based on the rules of individual applications. This architecture separates these layers and calls for developing procedures that guide developers to use the same reference models, tools and rules to ensure that the best ideas are carried throughout the organization. This architecture, coupled with proactive data management practices, will provide the following elements of the future environment: 1) seamless processing, 2) streamlined system development, 3) improved process automation, 4) improved data access and availability, 5) standards based data, and 6) proactive data management.

Seamless Processing

Across government, the age and complexity of existing information technology systems constitute a significant challenge. Historically, SSA’s programmatic business applications were developed at different times, for different purposes and were implemented as independent systems. SSA operations increasingly require employees, partners and citizens to make decisions and support processes that span the traditional SSA business lines. The concept of “Seamless Processing” is the ability to access data from various sources during a collection, access, or update event without having to log into and out of successive applications. Data will be accessible without regard to the data source, its native format or access mechanism. From the user’s perspective, all interactions will happen at the view layer in a single stream of activity. This experience is “seamless” since it presents no unnecessary interface, nor is there any need to capture and convert the data from multiple sources to serve the intended purpose. Therefore, internal and external users will find it simpler to access SSA services by using a single seamless data collection system. Eliminating the need to move in and out of multiple systems will streamline the process, resulting in increased efficiency and accuracy.

Streamlined Systems Development

The ability for developers to respond in short time windows is often characterized as "agile" development. For the purpose of this strategy, "agile" development means being able to shorten the time needed for definition of requirements, design, development and testing. In the current environment, data standards are often seen as impediments to progress since adherence to these standards takes time and are perceived to provide little benefit. As a result, system developers often prefer to begin development using ad hoc data structures that support their immediate need, but do not take into consideration the full data requirements of the project. This practice ultimately lengthens development timeframes when data issues identified late in the process result in costly scheduling delays to retrofit changes. In the future, an up-to-date Enterprise Data Model, a robust metadata repository and enterprise Global Reference Tables will foster agile development processes, reduce development time frames and improve quality, consistency and interoperability.

Improved Process Automation

SSA faces the dual challenge of rapidly rising workloads and increased program complexity. SSA cannot rely solely on staffing increases to meet these new demands and must leverage technology and automation where possible. It is important be able to rapidly develop and deploy new systems that will improve process automation in a timely fashion. The adoption of data-facilitated systems development processes will accelerate systems development, lower the total cost of ownership and, by improving the quality of business data, increase workload automation.

Standards Based Data

SSA data will be governed by a series of standards derived from a comprehensive Enterprise Relationship Model (ERM). The ERM will be the basis for standardizing data starting at the conceptual level and provide a framework for development of uniform standards at the logical (database relational models/entity relationship model) and physical (data values and formats) levels. Consistent standards at this level will promote interoperability, simplify cross-system business intelligence, facilitate data exchanges and lower operational costs. It will also accelerate the system design, development, testing and validation phases of the systems development life cycle (SDLC).

Proactive Data Management

Proactive data management is defined as an inclusive, coordinated action throughout the agency of anticipating future needs and building data-centric capabilities, so they are available when needed. Proactive data management is an inherently client-service oriented approach to managing data and the standards that govern it. It anticipates future needs, evaluates emerging capabilities and prepares the data environment in the advance of needs. Proactive data management includes: 1) codifying standard definitions for all data and 2) reaching consensus on common (shared) data definitions and data elements in order to present 'One SSA' to the client regardless of where their need exists.

Goals & Objectives of the Data & Database Strategy

As the agency's new data strategy is deployed, these goals and measures will assure technical initiatives do not inadvertently stray from their objectives. They will also provide a means to communicate value to stakeholders inside and outside of the agency.

Goal 1: Preserve and Protect SSA's National Asset & Preserve Public Trust - From Social Security's inception, the privacy and security of its data has been a high priority. It was a topic of the SSB's first regulation, published in 1937. Today, Social Security continues to maintain this focus, employing a multi-tier approach to ensure the privacy and security of its data. Over the past 72 years, the size and importance of SSA's data holdings have dramatically increased and now can rightfully be considered a national treasure. Accordingly, the agency's data holdings need to be protected from all manner of harm. For years, SSA has had a Disaster Recovery Plan in place to ensure there will be no loss of data in the event of a disaster. The agency has also recently opened a second data center to reduce the likelihood of an extended service interruption in the event of a disaster.

Preserving and protecting the agency's data holdings is not limited to preventing unauthorized access to, or destruction of key data. It also includes ensuring the quality and completeness of this data in order to provide accurate and timely service to the American public. Automated workload processing is dependent on quality data. Missing and invalid data can result in the need for manual processing, increasing costs and processing time and introducing the opportunity for error. Data integrity and quality are also key to data interoperability across the agency's programs. Appropriate disaster recovery processes, data quality, security, and privacy are all essential to the proper and efficient execution of SSA's mission and ultimately the public's trust in the Social Security program.

Goal 2: Expand Data Access and Availability - The expanding mission of SSA, the focus on Internet services, and the needs of SSA staff require that data be readily accessible by authorized users within their defined roles and available over the largest time period. If data systems go down and they will, restoration of data access must be prompt and, ideally, without impact to the user. As SSA continues to evaluate its standards for increasing systems availability, it is important to base future availability plans on thorough analysis of SSA business needs, technology alternatives, and projected costs.

Goal 3: Improve Technical Responsiveness and Agility - In the face of rapidly expanding workloads, SSA's systems must be easy to use across multiple sources of data and multiple technologies. Developers must be able to build and deploy new technical solutions quickly to meet new requirements. The core tools, data reference models, and system development lifecycles should empower development staff by providing them with easy to access technical capabilities that accelerate system development without compromising standards or quality. Data, data services, and technologies should be provided as standardized sets of capabilities that are governed by the enterprise.

Goal 4: Sustain Modernization - The best technical strategies are timeless, not because they anticipate future technical approaches with uncanny prescience, but rather because they conclude that they cannot anticipate future specific technologies. Change is the only predictable aspect of the future, and thus the strategy must position SSA to both live with and exploit changes as they emerge. An essential element to this capability is the notion of proactive data management. Proactive data management involves prioritizing data initiatives to prepare for new missions well in advance of when the data will be needed to support inevitably short development timelines. It begins with gaining a thorough understanding of the agency's data including business rules and other core standards that do not frequently change. This understanding of SSA's data must be documented and maintained with full participation of stakeholders across SSA and made available to all developers across the agency. This will require clearly understood "rules of the road," technical tools to enable consistent development and clearly communicated standards. It is important that across the enterprise, those who need these tools and standards have access to them. With effective data and other technical standards in place, SSA will be better positioned to support rapid development cycles without the risk of compromising data quality and integrity.

Data Quality

One of the main responsibilities of enterprise data management is to provide quality data to the data consumers. Employing proactive data management practices and suitable support technologies will enable the protection of SSA's data through the rendering of data quality software to establish a data quality firewall. The concept of a data quality firewall is that data validation software is used to ensure data is compliant with the appropriate data quality standards *before* it is stored in a database. Putting into place an automated process to ensure data quality is enforced when a record is created will protect the integrity of SSA's valuable data holdings and avoid downstream processing problems caused by errant data.

Because SSA has a long history and large inventory of existing data, the agency also needs to also employ other means such as data profiling to ensure the quality of its data. As documented in industry standard practices, data inventory and profiling is a three step iterative process: 1) Data Discovery and Inventory, 2) Data Capture and Data Staging, and 3) Data Profiling. The outputs of the profiling stage include "quality metrics" and "target definitions." These metrics and definitions set the stage for data cleansing and conversion.

Data Discovery and Inventory - In this initial step, information about the current system environment such as data dictionaries, business rules, naming conventions, existing metadata, data volumes, load strategies, and data architectural models is gathered and clarified with system stakeholders. After this information is identified and all systems are catalogued, an inventory checklist is created, maintained and monitored. Since this step will take a substantial amount of time to complete, tracking its progress by reporting the percent complete and reporting the gaps on a regular reporting basis will ensure progress and consistency for the team conducting the inventory. Periodic progress reports will create an awareness of systems and platforms across the organization for leadership, data owners, policy stakeholders, and business analysts alike.

Data Capture and Data Staging - The process of data capture and staging is performed to avoid any impact on the production databases during data profiling. The legacy data is in MADAM, IDMS, and VSAM files, and the target is a relational database architecture on a DB2 platform. Note, however, that the DB2 database is not immune to quality problems and that some data quality issues may also exist there. Nonetheless, data should be acquired in a staging area in order to conduct data profiling, data conversion, transformation and loading. In the case of existing modern database systems, an instance should be created for data profiling, so the production instances are not disturbed. Next, source to target mappings of the selected source data systems should be conducted. Then, the data can be extracted and loaded into the staging area. An industry and government practice is to capture the source data as-is with minimal data conversion, so the captured data can be analyzed and tested for conversion in the staging area without disturbing the source instances. However, in the case of SSA, the legacy source data is in mainframe files and the data acquisition is in the DB2 staging area and thus requires a thorough mapping of the data types, formats, and sizes.

Data Profiling - This step involves the activities required to identify and resolve data quality issues in the legacy sources selected for conversion during the mapping stage. The mapping baseline is an input to data profiling activities. It identifies the source systems and the data elements to be extracted and profiled. Conducting data profiling requires physical access to the source legacy data files and includes the following central activities:

- Loading legacy data
- Profiling data
- Resolving data quality issues

Data Cleansing - Quality issues found in the source systems during the data profiling stage, along with the decisions about how to remediate those issues, are inputs to the data cleansing phase. Data cleansing requires the physical changing of data values in the database and thus must be done with the utmost caution.

A number of basic steps are conducted to properly perform data cleansing:

- Define cleansing rules
- Develop cleansing scripts
- Run cleansing scripts to rectify data quality issues
- Stage the cleansed data for access

In order to make changes to the data, cleansing rules should be defined that guide the automated corrections. These rules should be written in pseudo code to ensure the business logic is clearly understood by the programmer responsible for developing the cleansing scripts. With clear logic about the conditions in which data should be changed and an understanding of what the data should be changed to, programmers can develop and run the cleansing scripts to officially correct the discovered data quality issues.

Maintaining Data Quality

Maintaining quality data within SSA's many repositories requires both systematic and programmatic efforts. From a systematic perspective, it is important to enforce data quality at the *point of data collection* to prevent the creation of new data quality issues. By standardizing data collection techniques through the implementation of ERM-based data services, SSA can ensure that the data collected from internal and external sources is accurate, complete and consistent.

From a programmatic perspective, it is important to conduct periodic data quality analyses in which data inventory and profiling activities are conducted to evaluate and cleanse the data that already exists. In addition, as proactive data management activities anticipate new data requirements, a quality control pre-requisite analysis should be performed to determine the appropriate parameters for the data need. This will help to ensure from the very beginning that only valid data is collected for these new fields.

Examples of SSA's Business Applications with a Focus on Data

This paper will close with a brief look at two of SSA's important business applications that are focusing on overcoming data related issues. The first is a system that verifies the name and Date of Birth of the person to whom a specific Social Security Number was issued. The second is a complex series of Business Applications that process SSA's most labor intensive and therefore most costly workload, disability claim adjudication. This application also is dependent on both structured and unstructured data, adding to its complexity.

SSN Verification

In 1935, the passing of the Social Security Act created a need to uniquely identify each person participating in the new Social Security System. This led to the creation of the SSN, a nine-digit number, configured as 999-99-9999. This number was intended to serve only for Social Security use. In 1936, Social Security began issuing a Social Security Card containing a unique SSN to each individual that would be working in a job covered by the Social Security program and for many years Social Security cards were printed with the legend "Not for Identification."

As time passed, use of the SSN expanded to purposes beyond those for which it was originally intended. All branches of U.S. Armed Forces adopted the SSN as their official identifier in the late 1960's and early 1970's. While the U.S. Internal Revenue Service (IRS) had always used the SSN to record Social Security contributions, by 1990, IRS also required children one year old to have an SSN in order to be claimed as a dependent on a tax return. Today, most financial transactions in the U.S. require an SSN. This expanded use of the SSN to all manner of uses has spawned a major workload for SSA -- that of verifying a particular SSN was issued to a particular person.

The verification process is initiated when an authorized organization sends an electronic request to SSA to verify the relationship between an individual and a specific SSN. The request must contain the individual's name and the SSN. The date of birth is also provided in some verification requests. When the request is received at SSA, an automated program determines if the submitted information matches SSA's records. SSA responds to the requester with an indication that the information did, or did not match the agency's database; certain requesters also receive additional information, such as death or citizenship data, a different possible SSN (if the submitted SSN does not produce a match) and or the reason for a mismatch. SSA's SSN verification process does not verify identity. There is no way for SSA to determine whether the person who supplied that set of information to the requesting organization is actually the person he or she alleges to be. SSA can only verify that the submitted information are related and match our records.

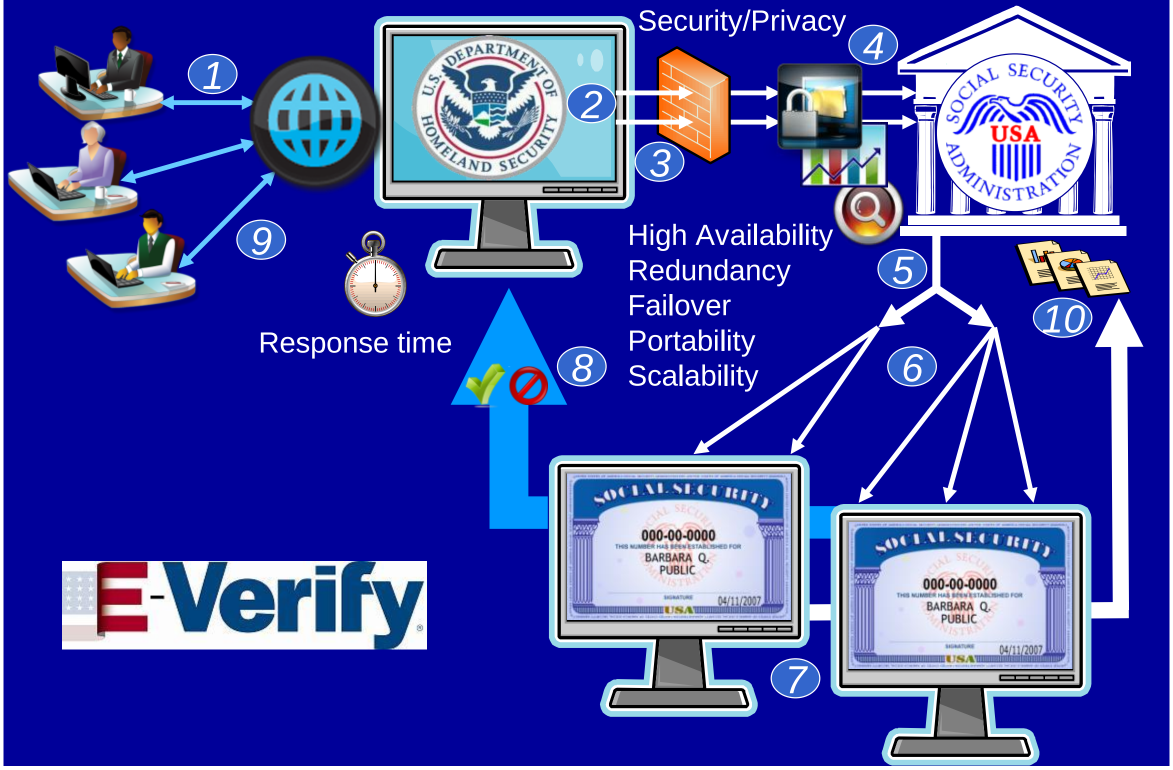
If the supplied information does not match SSA's records, it does not necessarily mean the information provided is wrong. The primary SSA resource for verification matching is the agency's Numident database. The Numident information, such as name, date of birth and citizenship is collected when a SSN is issued. The "number holder" should update this information by notifying SSA if they change their name or citizenship status. Failure to do so could cause a verification no-match, potentially causing an unnecessary delay, for example, in registering to vote.

Verifications can be requested for a variety of reasons and not all verification requests require the same level of precision. Accordingly, multiple matching routines have been constructed to accommodate this desired variability in matching tolerances. Some verification routines may require an exact match on all information supplied. Others may allow for transposition of digits of the SSN or reversing the name order. These different verification tolerances therefore make it possible to pass one verification process while failing another. The determination for which matching routine and tolerances should be used for each verification process is based on the purpose of the verification, business rules, authorizing legislation, and disclosure policy and law. For example:

> **The purpose of the verification -** The Help America Vote Verification system is an example of one of SSA's least rigorous verifications and is used only as secondary evidence to expedite voter registration, while E-Verify is more rigorous because it involves citizenship information and is essential to verify work eligibility.
>
> **Business rules -** When a new verification is requested, business rules dictate whether an existing verification routine is acceptable or whether a new verification routine must be developed.
>
> **Legislation -** Tolerances and/or specific guidelines can be included in legislation. Different legislation introduced at different times may result in mandatory variations in verification tolerances.
>
> **Disclosure policy -** Release of any information must be approved by disclosure policy and must follow privacy laws. Based on the reason for the verification, different disclosure policies and laws may apply.
>
> **Verification no-matches -** A "no-match" response is appropriate when the data submitted on the verification request is not consistent with data in SSA records. Some examples when a requester may receive a "no-match" response include incorrect or incomplete data in the request, typographical errors (including transposed digits), use of nicknames, or outdated data on SSA records (i.e. the number holder has not reported changes in information to SSA).

The importance of this unique and increasingly vital, national resource requires a suitable technical infrastructure to support its performance, protect its contents and maintain its access and availability. As a result, SSA has defined stringent technical requirements for its future verification processing environments. The diagram below depicts SSA's E-Verify architecture, which was designed to operate in an isolated processing environment with multiple failover capabilities, a higher level of redundancy, increased portability and scalability, and higher availability.

E-Verify Process and SSA's E-Verify Environment



E-Verify Process and SSA's E-Verify Environment:

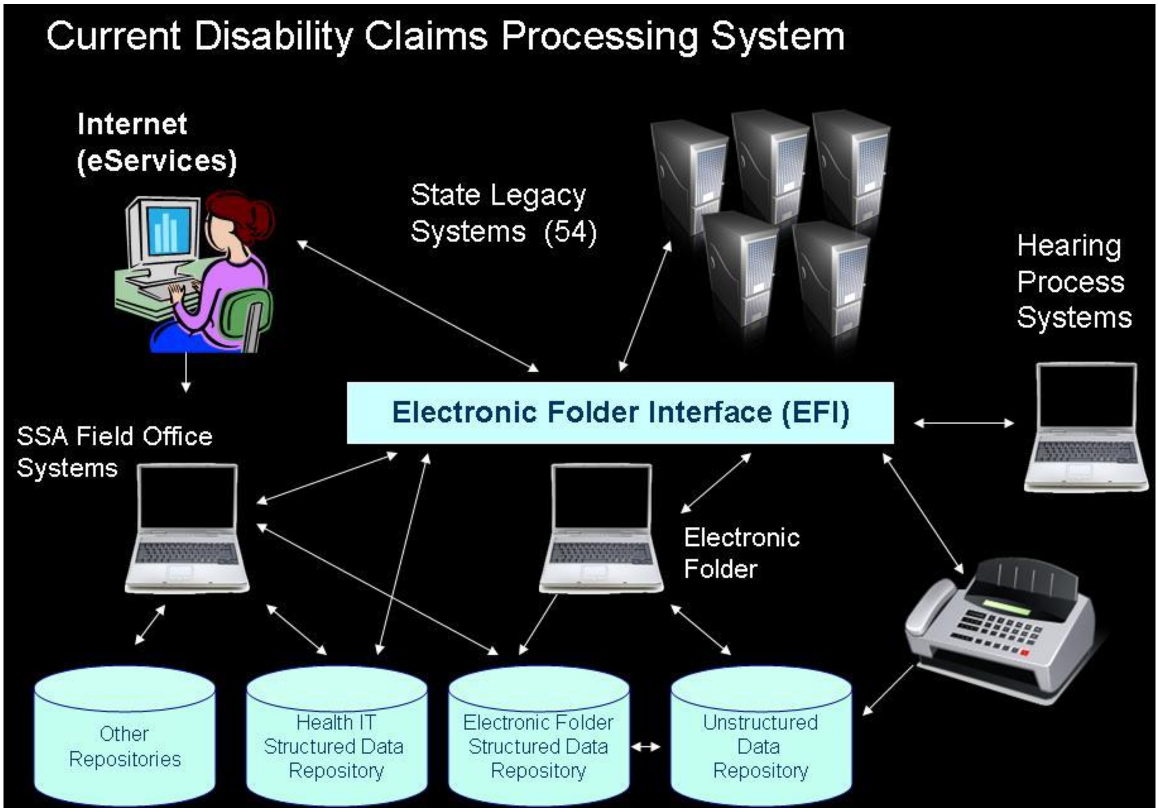
1. The employer enters new employee into the Department of Homeland Security (DHS) E-Verify system via a web connection.
2. DHS provides SSA with the employee name, date of birth, and SSN via a secured connection.
3. Data enters the SSA E-Verify system via a firewall, which provides data security and privacy
4. Isolation combined with a redundant architecture allows for increased availability. DHS workloads are not affected by other SSA work and SSA's programmatic processing is protected. Monitoring tools keep track of transaction volumes and the processing time.
5. Load balancing equipment distributes transactions equally between processing environments.
6. Multiple copies of the program exist in each environment to allow for failover and protect against system outages.
7. Data is processed through SSA's E-Verify program against one of two copies of SSA's SSN Master File (Numident).
8. SSA provides a response back to DHS.
9. DHS receives the results from SSA and makes a final determination on work eligibility before sending a response back to the employer.
10. SSA tracks transaction volumes and responses for management information purposes (e.g., systems monitoring, program evaluation, cost reimbursement).

Challenges in the Disability Claims Process

Social Security's Disability Insurance program is one of the most important safety nets for working Americans. It provides long-term payments to workers and their dependents in the event of a disability that prevents them from working. Although claims for disability benefits (approximately 4.5 million annually) do not represent SSA's largest workload by volume, the agency dedicates almost half of its administrative budget to the processing of disability claims.

By its nature, the process of determining if a worker is disabled and cannot work is labor intensive, involving the review of a claimant's medical records and, in some instances, sending a claimant to a specialist for a consultative exam. If a claimant is found not to be disabled they can request a reconsideration of that decision. If the claimant is not satisfied with the reconsideration finding, there are several additional levels of appeal. Although this process is federally funded, the original disability determination and the reconsideration processes are conducted by state employees in 54 state operated Disability Determination Services (DDS). The subsequent levels of appeal involve a face-to-face hearing with an Administrative Law Judge and are the responsibility of SSA's Office of Disability Adjudication and Review (ODAR).

Although processing disability claims was originally a completely paper-based process, great strides have been made in the past eight years and the process is now, by and large, electronic. The diagram below depicts the current process which begins when a worker files a claim for disability benefits. This can be done over the Internet, by phone or face-to-face at one of approximately 1,300 Social Security community-based offices.


Current Disability Claims Processing System
Internet (eServices)
State Legacy Systems (54)
Hearing Process Systems
Electronic Folder Interface (EFI)
SSA Field Office Systems
Electronic Folder
Other Repositories
Health IT Structured Data Repository
Electronic Folder Structured Data Repository
Unstructured Data Repository

The claimant is asked to provide information regarding his or her medical condition(s), the effect on their ability to work, what prescription drugs they take, and the names and addresses of medical providers (doctors, clinics, hospitals, etc) that have treated them for their condition(s). This information is captured electronically and placed in the Structured Data Repository (SDR). Jurisdiction is then passed to one of the 54 DDSs. There, a Disability Examiner, using an automated system to access the information provided by the claimant when the claim was filed, requests supporting medical evidence from the claimant's medical providers.

If the requested medical records are received in an electronic format, they are stored in the agency's enterprise repository for unstructured data, the DMA, as Tagged Image File Format (TIFF) images, where they are made available to Disability Examiners. If the requested medical records are paper-based, the medical records are routed to an out-sourced document capture center where they are scanned and converted into TIFF images to prepare them for entry into the DMA.

The structured data stored in the SDR and the unstructured data stored in the DMA are viewable as a virtual "electronic folder" and help form the basis for the disability determinations in the DDS and ODAR.

Two disability initiatives have recently been implemented and are designed to improve the speed and quality of the disability process. The first initiative, known as Quick Disability Determination, uses predictive models to identify disability claims that have a high likelihood for allowance. The structured application data provided by the claimant is passed through a predictive model where factors such as alleged disability, work history, and number of medications are weighted and combined to produce a numeric score. This score allows disability adjudicators to prioritize and expedite workloads. The second initiative is Compassionate Allowances which uses text mining and text analytics to identify individuals who are clearly disabled by the nature of their disease or condition.

There are a number of major initiatives underway to further improve the efficiency of this process. Currently, each of the 54 DDSs uses a different legacy case processing system. Although all are based on one of five different systems, all are customized and no two are exactly alike. This lack of standardization increases cost and decreases efficiency. A project is under way to build a standard case processing system for deployment to all DDSs.

Another major initiative focuses on moving to data-based structured medical records as their use becomes more widespread in the U.S. SSA already has a pilot process in place to accommodate medical records consisting of structured data and SSA is an active participant in the planning process for President Obama's proposed Health Information Technology initiative to convert the U.S. medical system to standardized structured formats for medical records. The benefits to be derived from this conversion are far reaching and, for SSA, it holds the promise of streamlining the disability determination process through the use of specialized software that can partially automate the medical assessment process.