National Patient Information Reporting System: National Data Warehouse

Data Integrity Verification

Plan

Version 1.0

March 2008

Department of Health and Human Services

Indian Health Service

Office of Information Technology (OIT)

Contents

Version Control

Version	Date	Notes
1.0	March 2008	Initial version. FY08 Contract Deliverable (D1.21.5) Accepted March 20, 2007

Overview

The Indian Health Service National Data Warehouse (IHS NDW) is in the process of loading healthcare information from both government and commercial healthcare sources. This central database provides a historical repository of patient registration and encounter information dating back to October 2000.

Data integrity, as defined, means that

> *data have not been altered or destroyed in an unauthorized manner. Data integrity is both a security and quality principle that prevents information from being modified or otherwise corrupted, either maliciously or accidentally.*

This Data Integrity Verification plan outlines the process to ensure data integrity in the loading of data from multiple and varied operational systems to the NDW. The Data Integrity Verification Plan is comprised of the following:

1) **Ensure data integrity**

   Verify that the data received and processed in the NDW accurately reflects the storage of data transmitted from the source system. This includes random sampling of export files received by the Integration Engine, before any processing has been done, and performing multiple data check comparisons to the data after being processed into the NDW.

2) **Integrity Reporting**

   The outcome of the comparisons will provide summary information for confirmation of integrity processing and reporting against performance measurements in data integrity. Detailed information will also be reported for formal analysis on any issues that arise in this process.

Data Integrity Verification Strategy

The philosophy behind the Data Integrity Verification Plan is to ensure the accuracy of the data. This process will help guarantee that any reporting out of the NDW provides an accurate representation of the data relating to the healthcare being provided to Native American and Alaskan populations. All parties involved have an important role in this goal.

Source data files are accepted "as is"; specifically, the NDW

- Relies on the FACILITY to ensure the QUALITY of the data that is being transmitted.
- Verifies the INTEGRITY of the data after it is loaded; that is, the data received is represented accurately in the NDW.

The data integrity verification process can also provide information to possible issues in the "quality" of the received data.

The NDW data integrity verification strategy utilizes the following levels of testing:

- Source to Target Counts
- Source to Target Data Verification
- Column to Column Verification
- Transformation Verification
- Exception Processing
- Summary and Detailed Results Reporting

Data Integrity Verification Process

This plan outlines the general approach in implementing the data integrity verification process. The NPIRS internal System Analysis Group will be the key component in establishing the optimum approach in application development. The ultimate goal is to automate the integrity verification process, but interim development can institute manual processes to ensure near-term implementation.

The data integrity verification process consists of several steps, and the flow is outlined in the following figure.

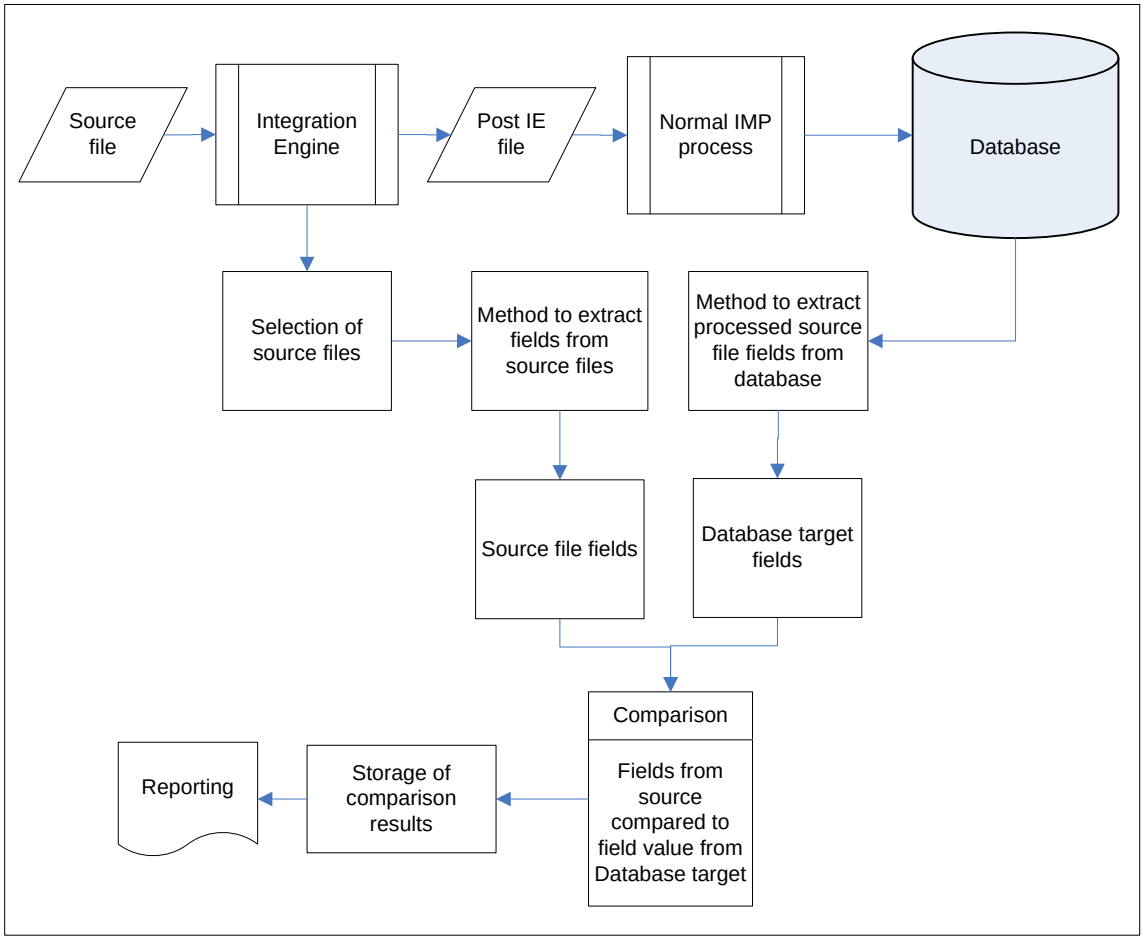

Figure 1.Data Integrity Verification Process

Source Files

A random sampling of source files sufficient to represent 10% of the data will be selected from sending sites that have sent exports to the NDW and which have been subsequently processed into the target tables of the NDW. A copy of these export files will be set aside and stored for integrity comparison. After the initial run, this process will run quarterly with a 10% sampling of files received since the last process was run.

Methodology will be developed to read these selected source files for comparison against the appropriate processed target Columns. The only transformation to these source Columns would be the character data type Columns that require transformation to a number data type for proper comparison.

Target Table Columns

A process will be developed to extract and read database target data, based on post processing of the corresponding selected source files. This methodology will take into account those transformation categories identified in the "Transformation Verification" section.

Comparison Process

The comparison component will match and compare the values Column-by-Column between the source file data elements and the target data elements. The process will extract and store any detailed discrepancy results for integrity validation summary and detail analysis and reporting needs.

Reporting

The process will produce three levels of information for Integrity Verification:

1. Source to target record counts
2. Summary reporting based on integrity errors and overall data elements evaluated
3. Detail information on each comparison for analysis and/or reporting

Column to Column Verification

This process deals with Columns that do go through any transformations. The value in the source file should match the corresponding target Column.

The Integrity Verification Process will:

- Verify the Column values in the source file to target.
- Ensure that the data mapping from the source file to the target is correct.
- Ensure that data received has been loaded accurately.

Source file data is accepted "as is;" that is,

- The NDW relies on the sending FACILITY to ensure the QUALITY of the data that is being transmitted.
- The NDW verifies the INTEGRITY of the data.

The Integrity Process, however, can provide information as to possible issues in the "quality" of the received data.

Transformation Verification

Minimal transformations to the data occur, based on established business rules within the source-to-target loading process. These transformations can be identified by the following categories:

- **Category 1:** Source Columns stored in designated target Columns as they are received AND transformed into proper formats in another represented target Column.

  Example: Dates - The text format date (_dc) is stored as it was received, and it is also transformed into a proper date (_dt) Column, yyyy-mm-dd.

- **Category 2:** Source data is stored as nulls in the designated target Column when data is received out of domain.

  Examples: Invalid numeric Columns; Character length out of bounds

- **Category 3:** Source data is transformed and stored in the designated target Columns as different values, according to established business rules and approved code sets.

  Examples: Gender (1 - Male, 2 - Female); Injury Place; Chart Number; Social Security Number

- **Category 4:** Some data types of the source Column are changed when inserted into the target Column.

  Example: Character to numeric

The Integrity Verification Process will validate the following categories of transformations:

- **Category 1:** The source text format date is stored as it was received in the target "_dc" Column. Essentially, this will be a Column-to-Column verification type. System function transformation into the "_dt" does not need to be verified.
- **Category 2:** These nulls will be validated as being properly set in the target Columns.

- **Category 3:** The different values and code sets conversions made in the target Columns will be taken into consideration for proper source-to-target comparisons.
- **Category 4:** The source Column data type will be converted to the target Column data type for proper integrity verification.

Reporting Results

Source file and target Column comparison results with be collected and stored on a cumulative basis to allow for analysis and reporting capabilities. The three main areas of reporting include:

1. Source-to-Target Count Reporting
2. Summary Reporting
3. Detail Reporting

Source-to-Target Count Reporting

In general, source-to-target count reporting verifies that the number of records received from the source system to the Integration Engine matches, within parameters, the number of records ultimately processed into the data warehouse.

The Data Integrity Verification process will rely on those fully successful loaded records from source to target. Processing business rules can “reject” specific records from being inserted into the appropriate NDW target table Columns. Therefore, the source-to-target counts would be adjusted to account for these expected differences.

A prime example of “rejected” records is encounters received not having their associated registration previously loaded in the NDW.

- The data is stored in the ERROR schema until the registration record is received and loaded.
- At the time of the source load, these encounters records are marked and counted as ‘rejected’, and the count value is stored in an ADMIN schema table..
- This ADMIN schema table can be referenced to adjust the overall Source to Target Counts verification.

Summary Reporting

The plan will provide summary numerical information on the records inserted into the database target tables correctly and incorrectly. A summary report will be generated indicating integrity performance metrics. These metrics will indicate the number of stored target errors compared to the number of source Columns evaluated.

Detail Reporting

Along with summary reports, detailed information will also be made available, primarily for analysis to correct system processing and improve integrity metrics. Detailed reports will be generated, allowing the capability to record these trends on integrity metrics. A secondary benefit to collecting this detailed information could to help identify and report on quality issues of source data.

Recommended Phased Implementation

The best approach to implement the data integrity verification plan will be to proceed in phases. This approach provides the ability to expedite implementation and apply any "lessons learned" to the follow-on phases.

The process is projected to be developed and implemented in the following phases:

Phase I

- Establish integrity verification on the HL7 structured files, which include source files from both RPMS and non-RPMS sending sites.
- Development within this phase:

  Step 1: Source data and target data will be verified on all Column-to-Column data elements.

  Step 2: Incorporate remaining source data elements that required "transformation" during target processing under established business rules.

Phase II

- Establish integrity verification on the non-HL7 structured files, which include source files from non-RPMS sending sites.
  - Simplified Format Encounters (SFE)
  - Simplified Format Registration (SFR)
- Source data and target data will be verified on all Column-to-Column and all "transformed" data elements.

Phase III

- Establish integrity verification on the Fiscal Intermediary (FI) structured files.
  - Contract Statistical Records (STATRECS)
  - Contract Dental Records (DENSTAT)
- Source data and target data will be verified on all Column-to-Column and all "transformed" data elements.

Phase IV

- Establish integrity verification on the RPMS CHS/MIS structured files (CHSSTAT).
- Source data and target data will be verified on all Column-to-Column and all "transformed" data elements.

Exception Processing

Source data may not be loaded to the target tables under certain circumstances. These circumstances, as discussed in the "Source-to-Target Count Reporting" section on page 6. These circumstances are routinely addressed to our customers, as follows:

- If an encounter cannot be linked to a registration record, it is stored in the ERROR schema tables until such time as the appropriate registration is received and the encounter can be promoted to the target tables. These records are reported by export via the Export Tracking web site and the Post Load report, and by area on the Promoter/Missing Registration report.
- Individual data Columns may be set to nulls when the data is invalid. This action and the associated data are tracked via a record in the ADMIN.LOAD_ERRORS table, and are reported by export via the Post Load report.

Known Issues

The Integrity Verification Process will highlight several previously identified issues that are documented in the Operational/Project Issues/Problems document. Currently, the resolution for these issues is scheduled into the Project Management Plan.

The known issues (Operational/Project Issues/Problems tracking #82) are:

- CPCS Remediation: Loading only one of potentially 25 occurrences of the following CHSSTAT Columns:
  - HCPCS_COST_AMT
  - REVENUE_CD
  - REVENUE_UNITS
  - REVENUE_COST_AMT
- PAY_DEST not loaded for CHSSTAT
- Authorization number (10 char) and PO_NBR (7 char) both mapping to PO_NBR
- CHS_COST_AMT and FI_CHARGED_AMT both contain Paid Amount
- DISCH_TP_CD is being loaded to DISCH_SVC_CD in FI files

These known issues will be noted and measured during integrity verification reporting, but will not contribute to the overall summary metric value.