

Appendix 5. Transformation Issues

Issue	Explanation
Location	Decide if extraction and transformation should take place on the source system or the SCS-DSS. The best solution may be to let the source system extract the data and let the SCS-DSS handle the transformation. SCS-DSS developers should try to minimize their system's impact on the source systems.
Updates ¹	Plan for updates. Updating data means moving a new batch of data from the source systems to the SCS-DSS. You must agree with your data source managers on how their systems identify new and revised records. This may be the most difficult issue to resolve in developing the SCS-DSS. Ideally, source systems track every addition and change with Date Created and Date Modified fields. Try to ensure that such a mechanism is detailed in each Memorandum of Understanding (MOU), Memorandum of Agreement (MOA), and contract; if not, updating your system could become problematic and resource-intensive.
Hardware and Software Differences	Hardware and software used by source systems may pose compatibility problems in moving data into the SCS-DSS. Identify and resolve these issues early in your development process. Source systems may be old and operating on mainframes using various generations of systems, programs, and files. If source systems perform the extraction and the SCS-DSS performs the transformation, you can minimize problems by ensuring sufficient documentation for incoming data and sufficient technical assistance for source systems and the SCS-DSS.
Record Structure	Constructing record structures suitable for your SCS-DSS is important when many of your sources are transactional databases. As part of the transformation function, you must reconstruct transactional records accurately in a structure that the analytical database will accept.
Data Format	Determine a set of data formats and implement them as you input the data into the Operational Data Store (ODS). With so many sources, information for the same data concept probably will be stored in several formats. Likely fields include dates and gender.
Data Cleansing	<p>"Clean the Data" using a two-step process:</p> <p>Step 1: Review records as they cross the boundary from the source systems into the SCS-DSS. Log and reject any records that do not conform to the business rules of the source system. Examples include eight-digit Social Security numbers, four-digit ZIP codes, and improperly formatted case IDs.</p> <p>Step 2: Inspect fields for consistency. For example, you may find variations in the spellings of cities, towns, and street names. Using the rules you established for the ETL process, the system will apply the correct format to all occurrences.</p>
Multiple Source Records	Decide which information takes precedence when conflicting data arrives from multiple sources. For example, based on a ZIP code from the source system, a city is identified as Springfield. A USPS database, accessed for the ZIP+4 value, calls it West Springfield. How does your system choose?
Sequencing	Decide on the order in which data tasks are completed. Sequencing applies to the steps within the ETL process for moving data into the data mart. Using the ZIP+4 example, you may decide to apply the USPS-generated city name before testing and correcting consistency. However, the generation of a ZIP+4 assumes you have a valid street address and valid ZIP code. If the street name is misspelled or misidentified (i.e., "Road" instead of "Avenue"), the process fails.
Summarization	If your SCS-DSS design includes automated data summarization processes, the system may perform these activities as data enters the ODS.
Logging	Set up an effective logging system. Your log will track records from sources not loaded into the SCS-DSS and tell you why they were rejected. This log is an essential element of the ETL process—it indicates instances in which the contents of CSES (and your other sources) do not match your SCS-DSS. It is important to reconcile these differences early in the development process to avoid future problems.

¹ A point of clarification: In the context of this book, the term iteration means a change or addition to the structure or operation of the DSS. Update refers to movement of data only.