



Practical data cleaning checklists for upstream and downstream teams

These checklists provide concrete actions for both data architects (upstream) and analysts (downstream) to prevent and address common data preparation pitfalls.

For data architects (upstream): Prevention checklist

Decision Point	Best Practice
Field storage	Store each piece of information in its own field. Never concatenate multiple data elements into a single string, even if it seems convenient for display.
File format	If fixed-width files are necessary upstream, also publish a delimited version for downstream users. Let one expert handle the import/export once rather than forcing hundreds of users to do it repeatedly.
Input Validation	Use controlled selection (dropdowns, radio buttons, etc.) for categorical data. Allow free text only when variation is genuinely needed. Ask: What kind of granularity supports downstream decisions?
Structured Data	For phone numbers, dates, IDs: collect components separately, strip formatting on entry, store in normalized structure. Concatenate for display but keep components available.
Date/time formats	Use unambiguous formats ("11-Sep-22 19:46:27" not "9/11/22 7:46:27"). Provide separate columns for common groupings: day, month, year, time. Anticipate downstream analytical needs.
Spreadsheet storage	Store data in column-centric structure with no merged cells and no missing values. Each row should contain complete information. Create presentation views separately if needed.

For analysts (downstream): Damage control checklist

Problem Encountered	Recovery Strategy
Concatenated fields	Use pattern-based text extraction with your tool's regex or parsing functions. Document your logic, validate against edge cases, and consider requesting properly separated fields for future data.
Fixed-width files	Invest time to import correctly once, then save as delimited format for future use. Share this converted version with colleagues. If this is recurring data, advocate for the upstream team to provide delimited version.
Inconsistent formats	For phone numbers, dates, or IDs: use find-and-replace with patterns to normalize. Create a new column with standardized format rather than overwriting originals. Build a reusable transformation script if this data updates regularly.
Messy categorical text	Create a mapping table from variations to standard values. Use recoding or lookup functions to apply consistently. Flag remaining ambiguous entries for manual review rather than guessing.
Merged cells/missing values	Use forward-fill functions to propagate values down through blank cells. Verify the fill logic matches the data structure. If structure is complex, consider going back to the source system before analysis.
Tool selection	If you regularly face complex data prep, evaluate whether your current tools support efficient transformation. Look for platforms that integrate data prep with analysis, provide visual feedback on data quality, and allow repeatable workflows.