

About Data Cleansing and GoldMine Data

According to Wikipedia, **Data Cleansing** means "**detecting and removing and/or correcting a database's dirty data (i.e., data that is incorrect, out-of-date, redundant, incomplete, or formatted incorrectly). The goal of data cleansing is not just to clean up the data in a database but also to bring consistency to different sets of data that have been merged from separate databases.**"

In the GoldMine world, this translates into three major topics:

1. **Merge-Purging Contacts** within the GoldMine database, to eliminate duplicate Main Contact records that may be part of the database. GoldBox offers a Merge-Purge that is commonly acknowledged to be the best available; significantly safer and more flexible than the one GoldMine offers. Please see my "**All About Matching...**" paper, on the web site's "**Getting**" **GoldBox** page, for important information about safely applying Merge-Purge.
2. **Creating and enforcing RULES** for the data in the fields of the various tables of the database; and **remedially applying those rules** via data alteration procedures, as required. For example, you may want to ensure that phone numbers for Contacts who are within the USA or Canada have GoldMine's USA phone format in all phone fields. Of course, your needs may be much more specific and sophisticated.

The details vary widely, depending on the fields involved; the rules you want to implement; and the sources of any outside "lookup data" that may be involved. GoldBox is equipped to handle virtually all these needs, and GoldMine can handle some of them. For example, while GoldMine's **Global Replace** tool can be helpful for some data cleaning operations, it is severely limited, compared to GoldBox. GoldMine can only handle very basic data cleansing in the Main Contact tables, while GoldBox can handle all the Tables of GoldMine. Also, GoldBox can accomplish **general outside-table-based replacements**, while GoldMine cannot. **GoldBox Views** can expand all these capabilities even further.

3. On an on-going basis, **managing any acquisition of data in such a way that the requirements of items 1 and 2 above are fulfilled automatically**, as part of the data acquisition process. For mass data acquisitions, this can be accomplished using **GoldBox Import/Updates**. And again, this is true whether the Import/Update is performed into the Main Contact database, or involves the other main data types of GoldMine, such as History, Pending, Details, etc. While this third item is not, strictly speaking, part of "data cleansing", it's vital in keeping the need for data cleansing to a minimum.

Now, let's take a closer look at the list of items in the Wikipedia definition that may need action on our part, to see how they correlate with the above list of three topics:

- "**incorrect, out-of-date**" relate to the most basic data characteristic of all: **accuracy**. This is simply about right or wrong, and is dealt with using the techniques in **item 2 above**. Some types of data can be checked against a standard of accuracy (for example, verifying City and State for the recorded Zip Code, against a Zip database that's known to be accurate). With the right tools (like GoldBox's Import/Update with Custom Matching), this can be a fairly simple process. Unfortunately, relatively few fields in most databases can benefit from this kind of treatment; still, it's worthwhile, where possible. Of course, it's important that the standard reference data used be **current**.

- “**redundant**” refers to at least a couple of things:

- a. **Poor database design**, so that the same information is stored in more than one place. Correcting this problem must start by eliminating all but one location for the data. Only then can we repopulate that location with the most reliable of the data that exist. GoldMine can often make such corrections to the Main Contact, while a specialized Global Replace or Import/Update (like GoldBox’s) would be needed for related tables. These techniques are included in [item 2 above](#).
- b. **Duplicate records (item 1 in our list above)**. In a relational database like GoldMine, this may mean duped Main records (i.e. two or more Main Contact records for the same person). It can also mean duplicates among the records in any of the relational tables. While duplicates among the related records may seem like more of a nuisance than a problem, dupes of the Main Contact records can create serious problems. It can result in a failure to find actionable data when needed; or worse, finding data that is not accurate.

GoldBox has **Merge-Purge** for removing duplicate Main Contacts; and **Dedupe one Table** for duplicate related records. As noted earlier, **do not rush into these procedures**; improperly done, they can damage your database so severely that only complete restoration from backup can put things right again.

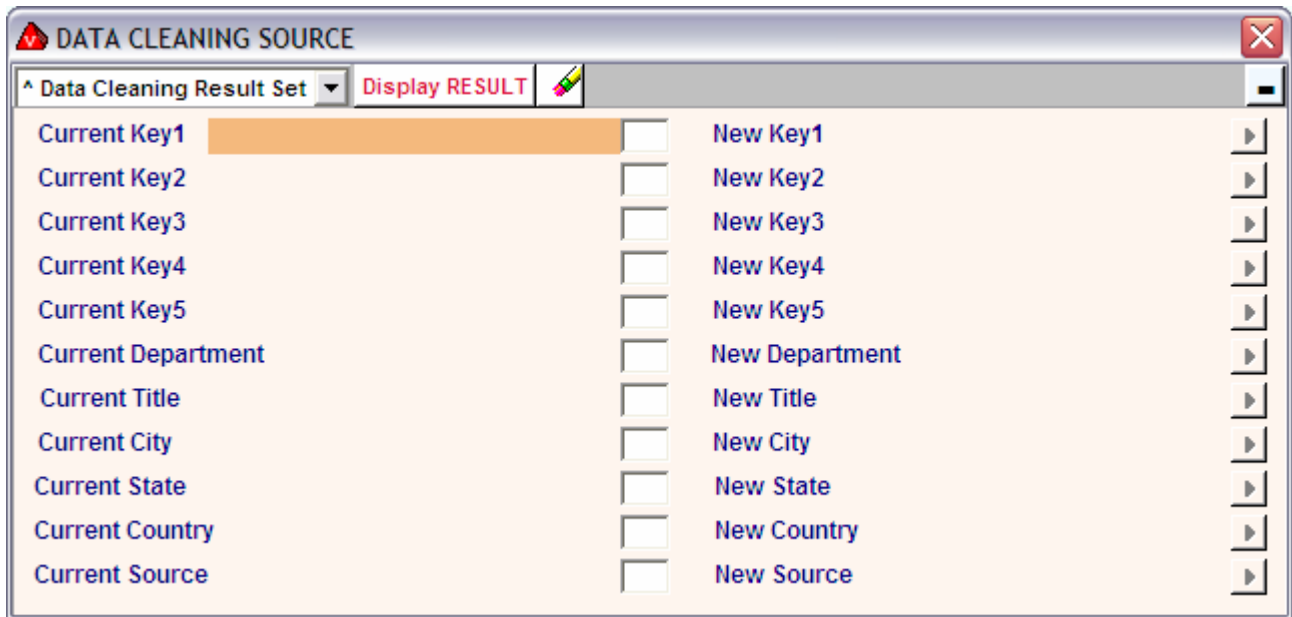
- “**incomplete**” refers to **missing data**, not data that has yet to be developed. Discovering records in which specific data should exist, but does not, generally requires a thorough knowledge of the way the database works, and the ability to write suitable queries or filters. Both GoldMine and GoldBox have tools that can help in this regard. This kind of work generally uses the tools discussed in [item 2 above](#).

- “**formatted incorrectly**” may refer to actual “square peg, round hole” data type mismatches. But much more often, it means that the data in a field **does not match any of the acceptable values in the Lookup List for the field**. This is referred to as **invalid data**. GoldMine does provide the option for the database designer to **force** manual data entry to be **limited to valid entries**; and **this option should be used universally**. But there are ways around that restriction, especially with Imports of new data, or Updates to existing data. GoldBox’s Import/Updates do provide ways to protect data fields from receiving invalid data via GoldBox Import ([item 3 in our list above](#)). GoldMine’s Imports do not.

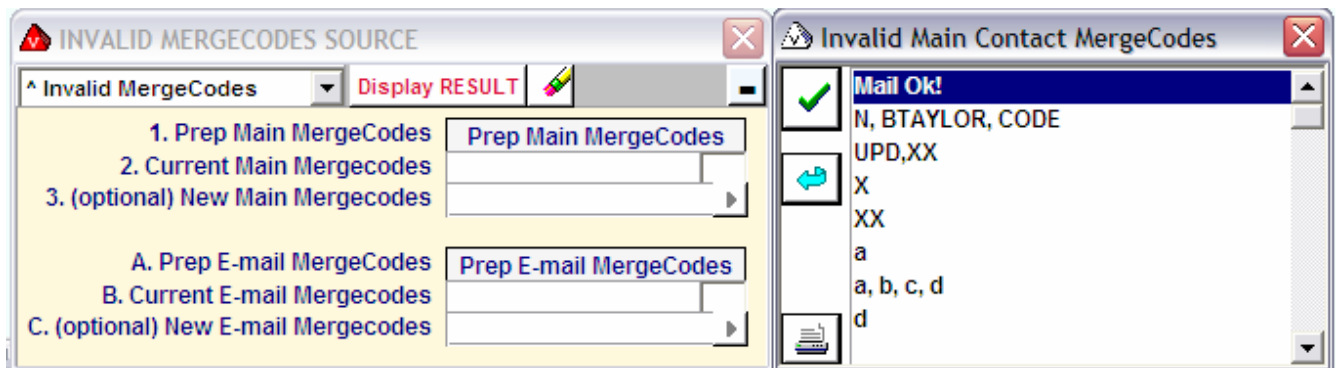
Why is validity so important? The ability to report on, and to analyze, your data are **profoundly affected by validity**. If you are tracking peoples’ ages using 5 age groups, then every record that has an actual **age**, instead of a valid **age group**, will present a problem. That data will drop out of normal queries and filters, and give you faulty analysis. The details will vary, but the same idea applies with any field that uses a Lookup List. How you **design your Lookup Lists**; and how you **maintain validity**; these are two of the **most important factors** in how well you can analyze and **use** the major investment that is your database.

For **correcting invalid data that already exists (item 2 in our list above)**, I have designed some special GoldBox Views. These take what might otherwise seem like a huge task, and trims it down to an astonishingly easy one.

These Views find invalid field data for you, and present you with a Lookup List of all the unique invalid entries for each field. You simply choose the invalid entry that you want to correct; then enter the valid entry to replace it. Push a couple of buttons, and GoldBox does the rest. There's even a version of it for multiple entry fields (like MergeCodes). It has an extra button, but otherwise works the same way. Here are screenshots of the Views that I have made part of my Ensemble package. (That means these Views are available to my clients for only the cost of the time it takes me to install them on their system...see the **Ensemble page** on my site).



Above, the invalid entries are revealed by the Lookup List for the left column; and the new values are selected from the (GoldMine) Lookup List in the right column. A second View opens when the **Display Result** button is pressed, and its Post button accomplishes the Update.



In the MergeCodes View above, the data is filtered and queried to produce a Lookup List of field entries in which **at least one** of the values in the single or comma-separated list is invalid. You then select the (optional) desired entry (as many values as desired). As long as all the values selected are valid, the View will kick off the replacement routine. If any selected value is invalid, the replacement will not begin (until the necessary correction is made).

The new value is shown as “optional” because you may not want to replace all the records that contain the selected invalid value with the same new value. In that case, the new value is left empty, and the **RESULT** View simply lists all the records that have the invalid value. You can then use the View to point to the correct GoldMine record, and make the desired changes one record at a time. This option also applies to the first View shown.

- There is one last item in the Wikipedia definition: “**also to bring consistency to different sets of data that have been merged from separate databases**”. There is a trend among users of GoldMine to consolidate databases; it’s a naturally attractive idea, because it eliminates the problems that come with having the same data in multiple locations. It is possible in part because of the greater capacity of SQL, and the management options provided by GoldMine’s Record Typing.

Typically, this is done by syncing the various databases into one. If the same GoldMine Accountnos are found in different databases, those records will automatically be combined by the sync process. But for other duplicates (i.e. dupes with different Accountnos), more is required. Typically, this is done by Merge-Purge ([item 1 in our list above](#)).

Consolidating databases is very tricky. I strongly encourage you to get expert assistance, if you are facing that kind of job. There are many, many things to consider before setting up such a procedure.

You probably knew that data cleansing is an important topic (why else read this paper?). But it may not have occurred to you that there would be quite so many aspects to it. And yet, it doesn’t really have to be terribly complicated.

Each database really is unique. Some issues will apply to yours, others will not. A bit of strategizing with an experienced consultant can save you a lot of grief and expense in the long run. Many data cleansing procedures can actually be automated, and run entirely on their own. Others will become minor chores to be performed a few times each year, once the initial cleanup is done.

The important thing is to make a start; you’ll be glad you did.