Suggestions and Guidelines for Improving the Quality of Your Specify 5 Data and Facilitating Their Conversion to Specify 6.

A. Bentley, Specify Software Project Staff

October 26, 2009

In preparation for your imminent conversion from Specify 5 to Specify 6 there are a number of things that you can do to ease the process and ensure that your Specify 6 database is as clean and free of errors as possible. Because a number of useful Specify 5 tools that can assist in this process (batch editing, batch re-identification, remove duplicates and remove unused records) are not yet implemented in Specify 6, it may be useful to accomplish some data clean up tasks before the conversion. Below is an outline of the procedures that will improve the quality and consistency of your existing Specify 5 data and facilitate the migration of your database to Specify 6.

1. **Geography Cleanup** –
   Because geography has been converted into a hierarchical tree structure in Specify 6, it is important to try and remove as much inconsistency and error as possible within geography data before conversion or the geography tree will be compromised. Cleaning up geography data consists of ensuring that all continents, countries, states and counties are spelt correctly and consistently used, e.g. “U.S.A.”, “Mississippi”, and “MS”. The batch editing tool can assist in converting unwanted variant entries. Create a search in Geography that contains these four fields, leave the criteria blank and sort by continent, then country, then state, then county. Export the results to Excel and print it out for easy reading and marking of those items that need attention. Now, using the batch editing tool and the SUBSTITUTE option (DO NOT use REPLACE unless you have a consistent list of items otherwise it will replace all entries at once!) you can replace inconsistent entries with the correct value.

2. **Taxonomy Cleanup** –
   There are two main areas that may require attention:
   a. **Misspelled taxon entry** – Edit the entry to the correct spelling. All associated collection objects will be changed automatically.
   b. **Two taxon entries for the same taxon** – This is a little bit trickier as you cannot create two versions of the same name in the tree (and batch edit does not work for taxonomy). The easiest workaround to resolve this problem is to search for the incorrectly named collection objects and change them all to the correct name, after which the incorrect name can be deleted from the taxon tree. Specify will not allow you to delete entries from the taxon tree that are being applied to collection objects. You can check to see which collection objects are associated with a name by right clicking on the name in the tree and choosing “Details”. This will invoke the taxon details form from where you can click on Related Records at the top and find which determinations are associated with that name. If none of the tables are active it means that the name is not being used and it can be deleted immediately. In some cases you may find that there are more collection objects associated with the incorrect version than the correct and it may be more time effective to change all the correct determinations to the incorrect determination and then rename the incorrect taxon entry in the tree.
   c. **Taxon Names with incorrect parents (misplaced entries)** – Use drag and drop in Specify 5 to move these taxon names into their correct parents.

3. **Agent Cleanup** –
   This is probably the most time consuming clean up task and you may decide that it is not worth the time required. You may have multiple agents for a single person e.g. John Public, John Q. Public, J.Q. Public, etc. The easiest method of removing these is to

make all versions of a single person's agents the same (all fields including address!) and then use the duplicate removal tool to remove the duplicate entries. With 1000's of agents in most collections, this can take some time but will make your database that much easier to search and the data more consistent.
(Note: Agent records with the same name but with different address details, are different Agent records.)

4. **Other fields** –
   The above holds true for other fields of data too (locality strings, text fields etc.). The more you can clean/correct these using the available Specify 5 tools the better your resulting Specify 6 database will be. But you will need to decide how much time to spend on fields of lesser importance.

5. **Pick lists** –
   All fields that are predefined as pick lists (not those created by you through the form customizer text field converter) should be checked for incorrect or misspelled entries. Incorrect entries can be edited through the form customizer while misspelled entries will need to be treated in much the same way as taxonomy entries. If the correct entry does not exist, simply edit the entry. If it does exist, all collection objects will need to be edited to the correct entry and then the incorrect entry deleted from the pick list. As in the taxon tree, entries cannot be deleted if they are being referenced by (applied to) any existing Collection Objects. (This can be done in Specify 6.x as well, if you do not want to clean up pick list data immediately.)

6. **Remove Duplicates** –
   This tool, along with the Remove Unused Record tool below, should be run as the final exercise before sending us your database and after all cleanup tasks are complete. This tool is found in the menu in Tools/Admin/Remove Duplicates and is run by table. Only certain tables can take advantage of this tool (Agent Address, Agent, Collecting Event, Geography and Localities) and the removal of entries requires that the entries are exact duplicates of each other (every field including fields that may no longer be active on the form but contain data). It is always good practice to start with those tables that are linked to others and then work upward, i.e. start with Geography, then Localities, then Collecting Events. As in the process of removing geography duplicates, you may be creating new locality duplicates etc.

7. **Remove Unused Records** –
   Another tool that can assist in your cleanup efforts is the Remove Unused Records tool. This tool (also found in Tools/Admin) will allow you to remove any records that are not being used by any higher related table i.e. Collecting Events that are not connected to any Collection Objects etc. In contrast to the Duplicate Removal tool, **the user can select which records to remove, and choose to keep valuable unused records** as there may be certain records that you wish to keep for various reasons. Again, only certain tables can take advantage of this tool (Collecting Event, Locality, Geography, Agent, Agent Address, Accession, References and Shipment). Select the table and an indication will be given of how many records are unused. You can then remove all or select individual records for removal. This will not only clean up your data but will also reduce the size of your database.

8. **New features in Specify 6 that may impact data in Specify 5** –
   a. Formatted catalog, accession, loan and gift numbers – Specify 6 now allows you to set a formatted string for these numbers to incorporate non-numeric components such as acronyms etc. **If you have such a requirement, please let us know as this cannot be altered once data is in place.** Once the database is created, the formatting is set. **[NOT SURE ABOUT THIS ONE OR IF THERE ARE ANY OTHER SUCH SCENARIOS]**

For more useful information you may want to read here –
-To see the schedule for database conversion or to see the status of your database conversion, go to (**insert conversion link here, if we are going to include it)**.
-For guidelines for importing External Data into Specify 6, go to http://specifysoftware.org/sites/default/files/data_import_guide.pdf. Written mainly for those intending to import data through the Specify 6 WorkBench this document also contains information about data clean up.

Once your database has been cleaned to your satisfaction (none of these are mandatory but remember that any errors/inconsistencies will be carried over to your new Specify 6 database) we can schedule you for conversion. You will need to detach your database using the detach option on the Specify 5 login screen and the "sa" username and password created during SQL server setup. **Important Note:** If you are working in a server environment this will need to be done on the server machine itself and not on your remote machine. The resulting detached *.mdf file can then be sent to us using our file dropbox at http://dropbox.yousendit.com/specify. The file size limit is 2 GB. If your .mdf file is larger than that you may need to zip it or we may need to seek alternative methods of receiving it. Please do not send us your file until you have been instructed to do so by us. You will be unable to work on your database until we have returned your new Specify 6 database to you and sending your file prematurely will result in an extended down time.

Other information that would be important for us to know at this time is whether there are any oddities to your data e.g. fields or tables that have been co-opted for purposes other than the intended use in the data model (besides generic text, number and yes/no fields which are intended for co-opting). Some of these may have been effected by us during your original conversion from a legacy system to Specify 5. Data and fields such as multiple catalog series, hostID usage, among others could have been used for your customized data.

If you have existing queries in Specify 5, these will be recreated in your new database. Some fields and some search criteria may not be included in the Specify 6 queries due to changes in the data model. Forms will need to be recreated using xml and we will try to match them as close to your existing forms as possible given the different limitations of xml (no more free positioning of fields – they are required to be in columns). If there are any changes that you wish implemented at this step, these will need to be articulated to us at time of conversion (ideally with screen shots and detailed information). This applies equally to data that you wish migrated into a different field than it is contained in Specify 5.

Any existing Specify 5 label or report formats will need to be recreated in iReports for Specify 6. We will need a list of reports that you use so that these can be recreated by the Project. We will attempt to match them as closely to your existing reports as is possible and if any changes are needed, they should be communicated to us at time of the database conversion.

Once your database has been converted to Specify 6, it will be returned to you for evaluation. You will need to have MySQL, Java and Specify 6 installed for this process – details of which can be obtained on our website at http://specifysoftware.org/content/download. An SQL data file will be returned to you, which will require that you run the Specify Wizard (http://specifysoftware.org/sites/default/files/specify_setup_wizard.pdf) to create a blank database into which you will restore your database. The values entered during the Wizard process are of little consequence as they will be overwritten by the values contained in the SQL data file delivered to you.

The evaluation process will involve you checking though queries and data entry) to make sure that the conversion placed all existing data into the desired fields, that no records are missing, and that forms and reports are to your satisfaction. We will provide a conversion report of data errors identified during the conversion. We will attempt to resolve problems (through

reconversion if necessary). After your review, we will ask you to acknowledge your acceptance of the conversion.

Remember that Specify 5 web interfaces, DiGIR provider interfaces, and publishing to GBIF will be interrupted with an upgrade from Specify 5 to 6.x. If it is absolutely critical that your existing network interfaces not be interrupted during the conversion process, you may want to continue to run your old Specify 5 database by re-attaching it to SQL Server. Any changes made to that old database however will not be reflected in your new Specify 6 database.

Please contact us by e-mail at specify@ku.edu, or by telephone 785-864-4400, if we can help clarify or provide other assistance related to the Specify 5 to 6 conversion process.