Importing External Data into Specify 6

This document describes options for importing collection data into a Specify 6 database and provides suggestions for preparing an existing DataSet for import.

A. Paths for Migrating Existing Collection Data to Specify 6

1. Conversion from Specify 5 to 6

The Specify Project has conversion software "The Specify Converter" which will do most of the processing to convert a Specify 5 database to a Specify 6 database. The ease and efficiency of the process depends on how consistently the Specify 5 data model has been used. If your Specify 5 database has data in fields that were co-opted from their original intent and scope, then the conversion process may require some additional manual conversion steps, because the field-to-field mapping between the two schemas will need to be customized. Also because of database design (schema) changes between Specify 5 and 6 are substantial for some data types (e.g. paleontological data), usage of some fields may necessitate some custom work.

If you are a registered, U.S., Specify 5 user in production and are ready to have your data converted to Specify 6, please contact the Specify Help Desk: specify@ku.edu or call +1 785-864-4400 to schedule a conversion. Functions have changed between Specify 5 and 6; it would be helpful to gain some familiarity with Specify 6, from a test installation or by reading the user documentation on the web site, before making the switch.

2. Migration to Specify 6 from another data management system

The Specify Project is able to convert data in legacy systems to the Specify 6 data model. To schedule a conversion from another database system, please contact the Specify Help Desk at: specify@ku.edu or call. We do not perform legacy data conversions for trial or evaluation purposes. Conversions to Specify are offered to U.S. research institutions.

3. Importing data in Excel spreadsheets into Specify 6 using the WorkBench

Although the WorkBench was designed as a tool for importing small sets of data - either from the field or other institutions, it can be used to import legacy data from an existing system given the limitations and guidelines outlined below.

4. Entering data into the Specify WorkBench for uploading into a Specify 6 database.

This pathway is an effective means for entering small DataSets from field notes or existing specimen labels for import into Specify 6. The WorkBench is designed to facilitate data entry and for many usages would be more efficient than entering new data directly into a Specify 6 Collection Object form.

Data keystroked or imported into the WorkBench from an external file are contained in Specify 'DataSets' which have a maximum size of 4,000 records. Entire DataSets can be committed to your Specify Database using the Specify Uploader tool in the Specify application. The Specify Uploader tool offers some options for comparing data in the WorkBench with existing data in the Specify database, but it is best to clean your data before using the Specify Uploader tool to avoid any record duplication. It is also best to check your data using the guidelines below to assure that it is correctly uploaded.

B. Data Cleanup: Consistency, Parsing, and Formatting

Effort spent standardizing existing data for consistency before importing them into Specify 6 will make a significant difference in the quality of the resulting Specify 6 database. The conversion process performs no consistency checking or correction to low quality data values except to validate that the values present are appropriate to the field type being used (date, number, text, logical, etc.)

Here are some suggestions for improving the quality of your collection data before converting them to Specify 6.

Correcting Typographic and Spelling Errors, and Improving Consistency

1. Ensure that common data values are spelled correctly. In some instances it may help to export data into Excel spreadsheets and sort columns to identify unusual values and facilitate cleanup. If you use Specify 5, one can take advantage of the batch edit tool to do some of this work. The conversion process will not correct misspellings in data field contents.

2. Ensure that frequently used data values are consistent. For example check that states, countries, agents, etc., are consistently spelled or abbreviated. Verify that text string fields, such as locality descriptions (Country, State, Province, City names, etc.) are standardized as much as possible. This consistency has implications for the usability of pick list data fields in Specify 6, where standard and consistent list values are important. Consistency also improves search accuracy, and for those data with tree-displays (Taxon, Geography, Location, Chronostratigraphy and Lithostratigraphy), removing data spelling, format and abbreviation inconsistencies will keep tree displays clean of unwanted branches and values.

If you currently use Specify 5, the batch editing tool can assist with these tasks. From within the WorkBench, the find and replace option is helpful.

Splitting Data Fields and Making Values Consistent in Fields Concatenated for Display

1. In order to move your existing data into the highly-resolved Specify 6 database design and capture all of their value, the contents of some data fields should be split (parsed) into individual data elements before moving data into Specify 6. For example, if your existing database stores the names of people (Agents) such as those of collectors, catalogers, and borrowers in a single field, if possible, you should split those names into separate first, middle, and last name fields so that the name elements will map precisely to the Specify 6 database structure. The Specify 6 database schema with all of the tables and fields identified is viewable on the Specify Project web site. Taking the time to do this will facilitate more precise searches of your collection and will help ensure new data consistency going forward.

2. Another example of data fields which benefit greatly from some pre-import clean up are those fields that are commonly concatenated together to display a long text string in a Specify data form. Location data are often formatted this way on forms (and in reports and labels) and typically include fields such as: Continent, Country, State or Province, County or Canton and nearest named place. Taxon data are also commonly used as a formatted string of values, each value representing a rank in the classification hierarchy. Depending on how you configure your Specify data forms, these strings can be used for queries and for picking values for new records. The level of consistency of the individual elements of these formatted text strings will impact the level of usability and efficiency of some data forms.

Complying with the Syntax and Value Requirements of Formatted Fields

1. Date Fields

Dates will upload into Specify 6 based on the format choice set in the Specify System Preferences menu.

Specify 6 allows for recording of partial dates, a common characteristic of historic collections. Legal partial dates may be missing a day, or a day and a month, but not a year. Missing values which comprise a partial date need to be formatted in certain ways for the Specify 6 to handle them correctly. This section describes how Specify recognizes partial dates, and the various allowable date formats which can be uploaded into Specify. The following date fields in Specify 6 are designed to handle partial dates: Collecting Event Start Date, Collecting Event End Date, Cataloged Date, Determination Date and Preparation Date). For all other Specify date fields, a full date is required; entry of partial dates will produce an error.

For uploading through the Specify Converter or thorough the WorkBench, and for Specify installations on U.S. date format computers, dates are required to be in one of two orders, either (1) month/day/year order, formatted as: mm/dd/yy or mm/dd/yyyy, or (2) dates may be in year/month/day order formatted as yy/mm/dd or yyyy/mm/dd. 00 must be entered for day and/or month in partial dates. However, a missing year (00/00/0000) is not accepted. One

other allowable format for U.S. date format operating systems is: dd/Mon/yyyy, where "Mon" represents the first three letters of the English month abbreviation (Jan-Feb-Mar-Apr-May-Jun-Jul-Aug-Sep-Oct-Nov-Dec). When using the three letter month abbreviation format, there are some additional constraints: the year must be represented with four numerals and only 00 is accepted for partial (missing) day formats. Only three hyphens '---' (ASCII 45, UTF8 45) are accepted as partial (missing) month values. Fully-spelled-out months are not supported for Specify WorkBench uploads.

On machines with operating systems that use European date formats, the uploadable dates must be organized in one of two formats: (1) dd/mm/yy and dd/mm/yyyy, or (2) yy/mm/dd and yyyy/mm/dd.

Days of the month may be represented as 1 or 01 through 31. Years may be formatted as 00 through 99, or as four digits: 0001 through 9999. Two year dates default to the 20th century, i.e. 19XX.

If a date field in Specify is configured to accept a partial date, the Specify Uploader allows two zeros (00) for unknown months or days. Also 5/00/YYYY is invalid and must be entered as 05/00/YYYY. A partial date with a year and day but no month is invalid and will produce an error.

Partial dates are displayed on Specify data forms (missing a day and/or a month) but are stored in the Specify database as the first day of the month and/or the first month for search purposes.

There are four valid date component separators: (1) a period or fullstop '.' (ASCII, UTF8 46), or (2) a forward slash or stroke '/' (ASCII, UTF 47), or (3) a space, or (4) a hyphen '-' (ASCII, UTF 45). Using other separators, such as dashes or backward slash will result in errors.

The following table shows examples of allowable date formats for Specify installations that are configured by the operating system to recognize U.S. date formats.

Date Content Template	Actual Uploaded Value and Format	Value of Date stored in Specify 6	Partial or Whole Date Displayed by Specify 6 (depending on date field format choice)
mm/dd/yyyy	00 00 1999	01/01/1999	1999
mm/dd/yyyy	00 11 1989	Invalid	Invalid
mm-dd-yyyy	11 31 1989	11-31-1989	11-31-1989
mm/dd/yy	00 00 04	01/01/2004	2004
mm/dd/yy	04 00 04	04/01/2004	04/2004
yyyy.mm.dd	1989.00.00	01.01.1989	1989
yyyy/mm/dd	1999/01/00	01/01/1999	01/1999
yyyy/mm/dd	1999 00 00	01/01/1999	1999
yyyy/mm/dd	1999 02 00	02/01/1999	02/1999
dd/Mon/yyyy	04 Jan 2004	01/04/2004	01/04/2004

dd/Mon/yyyy	00 --- 2004	01/01/2004	2004
dd/Mon/yyyy	04 --- 2004	Invalid	Invalid

2. Latitude and Longitude Values

Latitude and longitude values can be represented in one of three ways: (1) degrees minutes seconds, (2) degrees decimal minutes, or as (3) decimal degrees.

Uploaded Latitude and Longitude values without a pre-pended sign and without a post-pended compass direction are uploaded as positive numbers (indicating N Latitude or E Longitude) into Specify. A negative sign (S Latitude or W Longitude) must be a hyphen character '-' (ASCII, UTF 45). A dash character (UTF 8210, 8211 or ASCII 150, 151) will throw an upload error.

Degrees can be indicated by a degree symbol: ° (ASCII Dec 176/UTF 0176), or 'd' or a space or both. Minutes can be indicated by a single quote: ' (ASCII 39/UTF Dec 39) or a space or both. Seconds can be indicated by a double quote: " (ASCII 34/UTF 0022) or a space or both. 'm' and 's" are NOT recognized as indicators of minutes and seconds.

Values for Latitude and Longitude must be in one of these formats:

Degrees Minutes Seconds	Degrees Decimal Minutes	Decimal Degrees
- 32 45 16.8232	- 32 16.8232	16.8232
- 32d 45' 16.8232"	- 32° 16.82'	16.8232°
32d45'16.8232"	32°16.82	16.8232 N
32°45'16.8232"	32d 16.82	16.8232° N
32° 45' 16.82"	32 16.8232 N	
32 45 16.8232 N	32° 16.82' N	
32d 45' 16.8232" N	32°16.82 N	
32d45'16.8232" N	32d 16.82 N	
32°45'16.8232" N		
32° 45' 16.82" N		

See the Specify Help system for more information on storing and transforming latitude and longitude values in Specify.

3. Catalog, Accession, Loan, and Gift Number Field Formats

In Specify 6, specific formats can be defined for catalog and accession numbers as well as for loan and gift numbers.

- The formats of these fields will either match the format in your existing database, if you already have Specify 6 set up and running, or they will be set in the Specify Wizard when creating a blank Specify database for the first time. The data in the WorkBench must match the existing format in Specify in order to import successfully. In cases where

existing data in a legacy database is being uploaded by the Specify Project into a new Specify 6 database, these number fields will be formatted to match your existing legacy data syntax.

- If using the Specify WorkBench to import data into Specify 6, the Catalog Number format in the WorkBench DataSet must match the field format (including the length) for Catalog Number field in the Specify 6 database.

- If you have an existing Specify database with Catalog Numbers (or one of the other formatted number types) set to be numeric only AND auto-incrementing, the Catalog Number (or other) field can be left blank in the WorkBench. Catalog Numbers will be created automatically in Specify 6; the new numbers will begin their sequence based on the last number in the database.

- A Catalog Number field which is not formatted as an auto-incremented number in Specify 6 must contain data in the Catalog Number field of all records being uploaded (and they must be the same format).

- Also, if uploaded records contain catalog numbers, the specific values being uploaded must not already exist in your database. The same rule applies for Accession numbers when you are entering data to just the Accession table.

4. Tree-Structured Data

Specify 6 displays taxonomy, geography, biostratigraphy, lithostratigraphy and storage location information in tree formats, thus allowing easy browsing and editing of data values within the context of their place in a hierarchy and their parent/child relationships.

For uploading data to these fields, the importance of consistent, accurate data is paramount as incorrectly spelled items will create unwanted nodes in a tree. Any records being imported with incorrect relationships (e.g. a genus incorrectly positioned within the wrong family) will be placed as such in the Specify 6 Taxon Tree. The Specify Uploader has no intrinsic knowledge of taxonomy or of relationships in the other kinds of data trees.

There are three basic methods for populating tree-structured data within Specify 6:

1. One can load data from a predefined authority file. Specify makes use of the Catalog of Life and a freely available world geography data source to fill the taxon (optional) and geography trees. You can also supply a different authority of your choosing at the time of conversion if we are converting the data for you. Your data can then be matched up to this preexisting authority. If there are misspelled or incorrect elements they will not match the preexisting ones and new nodes will be created for these elements, which make them easy to find and clean-up in the tree display.

2. The trees can be built using your existing data during the conversion process by Specify staff or skilled IT personnel. This method will attempt to mimic your data structure for these elements and build the tree for you. If there are misspelled or incorrect elements they will be entered as such into the tree.

3. You can use the WorkBench to create these trees using Excel spreadsheets of data.

    - The trees in Specify 6 are configured by the user with ranks and each rank can be set to be required or not. 'Required' ranks are those that must have data values present in an uploaded record for the record to be uploaded successfully.

    - All ranks with data being uploaded in records from the WorkBench must be present in the respective Tree Definition in Specify.

    - When uploading only Taxon data (to add to the Taxon tree) use the mapping fields from the Taxon Only data types.

    - When uploading Taxon data together with other data, the Genus, Species, Subspecies and Variety fields need to be mapped to the Determinations data type.

    - If the highest rank in the DataSet does not match a rank in the taxon tree, new ranks will be created labeled with the name of the DataSet being uploaded.

5. Pick Lists

Data being entered through the WorkBench into a predefined pick list field in Specify 6 or data that you wish to have in a pick list will need to match the predefined terms in that pick list. For existing DataSets being converted to 6, pick lists can be made to match the data which is provided in most cases. In some cases there are very rigidly predefined pick lists of terms (type status etc.). In these cases users must ensure correct spelling of these terms. Values are also case sensitive.

When entering data through the WorkBench, terms must match predefined user and system pick lists in order to upload successfully. If there is a new term in a WorkBench DataSet, one must first modify the pick list in Specify 6 to include this term through the schema configuration tool and then upload the data.

6. Field Length and Data Type

Fields of data cannot exceed the field length of the mapped field as set in the Specify 6 schema. All additional data will be truncated and not imported. Field length can be viewed through the

Schema Configuration Tool or on the Specify Project web site. In general, string fields are limited to 32 characters, while number fields are restricted to 24 characters, but there are numerous exceptions and it is best to check. Number fields will not accept alpha characters.

7. Required Fields

Fields that are set in the data model or through the System Configuration tool to require a data value to be valid, must contain data, both in existing external data and in DataSets being uploaded through the WorkBench. Again, modifications can be made in the Schema Configuration tool to accommodate non-existent data, of the pre-set Specify requirement for data does not fit your data handling methods.

8. Preparation and Preparation Type

Specify 6 uses Preparations as the linking mechanism between Collection Objects and Interactions (loans, gifts etc.). In other words, in Specify 6, technically and precisely speaking, Preparations are loaned, not Collection Objects. If a Collection Object (i.e. a "lot" or "specimen") does not have any Preparations associated with it, then that Collection Object cannot be loaned or gifted within Specify. If Preparation and Preparation Type do not exist in your existing database, e.g. most herbarium collections do not by default recognize the concept of a "Preparation" (as they almost always have one type "an herbarium sheet" and one copy), you will need to create Preparation and Preparation type values either in the file you send for conversion or in your WorkBench DataSet in order to be able to make use of the Specify Interactions modules.

C. Specify WorkBench Requirements and Limitations

In order to ensure that data elements will be placed in the correct fields in the Specify 6 data model it is important to ensure correct mapping of these fields in the WorkBench mapping tool. In some cases this is obvious, but in other cases, especially for generic text, number and yes/no fields that may have been co-opted for a certain data element, it may require some inspection of the schema using Specify 6 Schema configuration tool to determine which is the correct field. By searching the relevant table in the Schema configuration tool you will be able to determine which field to use based on the caption assigned to that field. DataSet mappings can be reused from one DataSet to another to maintain consistency across files (especially important when working with files that are larger than 4,000 rows that will need to be split). In order to reuse mappings, the files need to contain the same columns and headers. If the files differ in this respect, the user will not be prompted to take advantage of this option.

Remember also, that data mapped to fields in the data model that do not exist on a default form will not automatically be visible. The form will need to be modified to include that field using the XML template for forms. The Specify team would be happy to help with any form customization

that is needed. Contact the Specify Help Desk: specify@ku.edu or call +1 785-864-4400 to arrange for this service.

Before uploading from a DataSet into a Specify 6 database, the WorkBench will validate the DataSet. This process will ensure that the necessary linking fields are present to link the various affected tables together, i.e. you will need a locality name field to link a locality to a Collecting Event etc. Some errors in the data format are also detected at this stage. All errors detected during this validation process are required to be fixed before uploading will be allowed. The affected cells can be edited during this step by clicking on the error row in the Specify Uploader message window. Once the affected row has been corrected type the Enter key to move to the next affected row. Once all affected rows have been attended to, save the DataSet and the validation process will run again. Only when all affected rows have been corrected will the Upload button become active.

In order to successfully map the various one-to-many relationships within the database, certain fields have been "flattened". For instance; if you have multiple collectors you will need to map each person's first, last (and, if present, middle) names to an individual field. The WorkBench mapping tool provides this facility by having Collector Last name 1, Collector Last Name 2 etc. fields. You are limited to 8 collectors by the mapping tool. You are also limited to two determinations, four accession agents, 8 preparations and one reference work. You also cannot have missing collectors in the string i.e. Collector 1, Collector 2 and Collector 4 where Collector 3 is empty or missing.

The WorkBench data model is a scaled down version of the complete Specify 6 data model. In some cases, a field in your DataSet may map to a field in Specify that is not included in the WorkBench data model. In this case it would be better to contact Specify to allow us to map your data successfully for you – especially if you are trying to map legacy data from an existing database system. Contact the Specify HelpDesk at specify@ku.edu or call +1 785-864-4400 to schedule a conversion. Also, please let us know if there is a field that you would like to see in the WorkBench and we will be happy to consider adding it to a future release.

WorkBench DataSets are limited to about 4,000 rows, the exact number depends on the amount of information in each record and how much memory the Java environment has available. We recommend 4,000 as a target upper limit for most cases. However, multiple WorkBench DataSets can be uploaded in sequence to a Specify 6 database with the data field mapping easily applied to one DataSet being transferrable and applied to subsequent DataSet uploads.

The WorkBench also does not update or augment existing records (e.g. for adding georeferences to existing locality records). We are considering a batch update feature for a future release.

The Specify WorkBench will check the records being uploaded against existing records in the Specify 6 database to identify possible duplication. The Settings button in the Specify Uploader allows the user to identify the action to be taken if more than one matching record is encountered:

- **Prompt** will display the matching records in a dialog and require the user to choose which of the matching records to use. An option to add a new record is also provided.

- **Add New Record** will simply add a new record and not search for matches. This may create duplicate records in the database.

- **Pick First** will choose the first match found and not prompt the user. A message indicating that this action has been taken will appear in the message portion of the Specify Uploader.

- **Skip Row** will not upload rows that match multiple database records.

- **Remember Choices** will treat the first record with matches exactly as a 'Prompt' above, but will remember the choice and use it again when identical values are found for succeeding records.

- **Match Empty Cells** when set to true will require fields mapped to empty cells to be empty when searching for matches. It will disregard empty cells when set to false e.g. Given a dataset row:

  CollectorFirstName1 = "" (empty) and CollectorLastName1 = "Nixon"

  and a database that contains agent records:

  FirstName = "Angela", LastName = "Nixon" and
  FirstName = "Shemp", LastName = "Nixon".

  If MatchEmptyCells is false, then two matches are found and prompted for the row. If true, then no matches are found and the record is uploaded to the database.

Duplicate record checking is performed on a data table-by-table basis; it is important to understand how the mapping (or not) of data fields in the WorkBench to the Specify 6 schema will affect the disposition of duplicate data. During the upload process, the Specify Uploader tries to 're-use' existing records whenever possible. When matching data in the WorkBench to data in the Specify database, the Specify Uploader will only recognize a record in the WorkBench as being unique if one of the fields being uploaded contains unique information. Also, the fields must exist within the same table. For example, if this record exists within the Specify database:

Locality: Martha's farm
Country: USA

State: Kansas
County: Douglas
Lat: 34.1000
Long: -98.1000

and you are uploading the following from the WorkBench (Example 1):

Locality: Martha's farm
Country: USA
State: Kansas
County: Douglas

The Specify Uploader would interpret these as matching records, because each of the fields being uploaded contains matching data.

However, if the WorkBench record included (Example 2):

Locality: Martha's farm
Country: USA
State: Kansas
County: Douglas
LocalityText1: In the swamp.

Then the Specify Uploader would interpret this as a new record because new information was included within a field in the Locality table.

If the WorkBench upload included (Example 3):

FieldNumber: TGN 3452
Locality: Martha's farm
Country: USA
State: Kansas
County: Douglas

A new Locality record would not be created because the Field Number field is part of the Collecting Event table and not considered when matching records in the Locality table. The Field Number value would be entered into the Collecting Event table and associated with the existing Locality record.

Also, if the upload included (Example 4):

Locality: Martha's farm
Country: USA
State: Kansas
County: Douglas
Lat:

Long:

If 'Match Empty Cells' is turned on this would be considered a new record. If it is not turned on then a new record would not be added in the database.

As you can see, it is beneficial to understand both the records that exist in your database and the records in the WorkBench and how you wish for them to interact. Any errors detected during the upload process will result in that specific row being omitted from the upload (i.e. not uploaded). All errors will be shown in the message window of the Specify Uploader--the affected rows will need to be corrected and re-uploaded after the initial upload.

Once data has been uploaded, the process is not finalized and records are not copied to the database until the “Commit” button has been clicked. Before this step is completed the data can be validated by using any of the modules in Specify – view the data in a form, query or simple search for the data etc. This is always an important step to ensure that data has been mapped correctly and that the intended result has been achieved before committing the data to the database. If errors have been made, the upload process can be rolled back and undone to allow for re-mapping or data cleanup before attempting the upload again.

Don’t forget to press ‘Commit’ when you are ready to finalize an upload!