Specify Software Project
www.specifysoftware.org

Using the Specify Web Portal

Installing, Mapping and Loading Data, Use

Table of Contents

1. Web Portal Software Requirements

Note: Installing, configuring, and securing the software infrastructure required for a Specify Web Portal is a systems-level activity. The procedures described in the documents linked below should be undertaken by a server administrator.

The Specify Web Portal is designed to be installed on a Linux server computer with Apache Tomcat. The installation documentation for the Web Portal is based on Ubuntu Linux, with minor changes to the installation procedures other versions of Linux should also work. Specify Web Portal software is available from GitHub as noted below.

Software components to be installed:

- Apache Tomcat 7
    - Software and documentation are available from: http://tomcat.apache.org/download-70.cgi
    - Using network port 80 is recommended (Tomcat 7 default is 8080).
    - Solr is required by the Specify Web Portal and it is automatically included in the Tomcat 7 installation.

- Specify Web Portal Software
    - Obtain the Specify Web Portal software from: https://github.com/specify/webportal-installer. Installation instructions are available there.

- Specify Attachment Server (optional)
    - The Specify Attachment Server is only needed if viewing collection object image attachments in the Portal is desired. The Web Portal can be used without an attachment server but then images will not be associated. General information on the Attachment Server can be found at: http://specifysoftware.org/specify-6-5, and at http://specifysoftware.org/download.
    - Attachment Server Python software code and installation notes are available at: https://github.com/specify/web-asset-server.

2. <u>Overview: How Specify Organizes and Exports Data</u>

This section summarizes how to organize and export Specify data for external use. Identifying the Specify data fields to be exported (e.g. those desired for the web portal) is the first step. Then, an interactive task called 'schema mapping' or 'concept mapping', associates Specify data fields with data concepts in an **Export Data Map (EDM),** which defines which data fields and records should be put into an exported data file. Specify facilitates that task with a software module called the "**Schema Mapper**", which is accessible from the Specify menu bar under 'System'. The Schema Mapper has a visual, interactive window which lets you choose Specify data fields for export and map them one-at-a-time to data concepts in an Export Data Map.

Mapping data fields in Specify to data concepts in order to create an EDM can be done with or without reference to an external **Export Schema.** An Export Schema (ES) is a data file template produced by a separate project or agency which specifies how they expect data to be organized to be uploaded to their databases. There is no external Export Schema to produce an EDM for the Specify Web Portal, because the contents of that map are specific to each collection's web portal customization.

By using 'Operators' and 'Criteria' for Specify fields in the Schema Mapper, the contents of each data field to be exported can be filtered by the conditions you set and sorted, displayed, or hidden. One constraint which Specify itself places on the external file is that each exported record must be unique; we explain how to avoid exporting duplicate specimen records below.

After an Export Data Map has been created and saved in the Schema Mapper, the actual exportation is a two-step process with Specify. First, a separate program, the Specify **Data Exporter**, which is located in the Specify installation directory (../Specify/bin/DataExporter), takes a completed EDM and uses it to create a copy or 'cache' of your data. This copy is formatted as specified in the mapping, but the results are stored within your Specify database inside a single, flat, MySQL table. This internal copy of your data is a static snapshot of your database; it is only updated when the internal cache is re-built using the Data Exporter program. The Data Exporter shows the status of each cache or copy of your data which you have created; it indicates whether a copy is current or whether it needs to be updated because records in your database have been added, removed, or changed since the cache was built.

A second step copies your export data cache to an external file. Clicking on the 'Export' button in the Data Explorer creates a tab-delimited data file to be named and saved. Alternatively, using the 'Export to Web Portal' button creates an external .zip file which includes a directory of configuration files as well as your exported data for a Specify Web Portal.

3. <u>Create an Export Data Map</u>

An Export Data Map, in part, defines a query used to export data from Specify which are indexed and made searchable within the web portal. Data fields included in the Export Data Map are used in the Web Portal in three different ways—they can be configured as search fields listed in the "Advanced Search" window pane, they can also be selected to appear in the spreadsheet-like view which appears under the portal "Record" tab. Finally, all exported data fields appear in the floating pop-up window of the detailed data view (not customizable).

> ☛ *The primary goal when creating an Export Data Map is to consider which data fields you want to use in each of three web portal views—for the left-side query panel, specimen data spreadsheet view, and single specimen detail view--and to include all of them in the EDM. Data in Specify database fields not mapped to export schema concepts will not be available in the portal.*

The Schema Mapper is similar to the Specify Query Builder in that it lets you choose data fields and conditions for formulating a query (in this case to select and order data for export), but with the addition of the capability to map those data fields to an Export Data Map. The text of the labels in the Schema Concept column ultimately are used as the names of the data columns in the Web Portal.

Section 3.1 describes how to create a custom Export Data Map for a Specify Web Portal. Section 3.2 describes how to use an existing external schema to guide the mapping and development of an Export Data Map.

3.1 Define a New Export Data Map without Using an Export Schema

Log into Specify and open the Schema Mapper within the Specify application by choosing: *System > Schema Mapper*.

> ☛ **You must use a Specify user account at the Specify administrator-level to be able to run the Schema Mapper.**

One option for defining an Export Data Map is to choose *'no schema'* in the "Create New Mapping" dialog list when using the Schema Mapper. This choice is the needed if you are creating an Export Data Map for the first time in order to custom organize and export data to a Web Portal. The Specify software installation does not provide a default Export Data Map, it is created as part of the export data mapping process, and is custom for each collection and associated web portal interface.

The Schema Mapper opens with a list of Specify fields at the top of the work space and like the Query Builder enables you to navigate through the Specify database structure to identify and pick the data fields to be mapped and exported. By selecting a field from the Specify database and clicking the ✚ button, a data concept will be created with the same name as the Specify field. The schema concept name can then be changed by simply clicking in the text box and typing a new name for the concept. The concept names in the EDM become the data field column names in the Specify Web Portal.
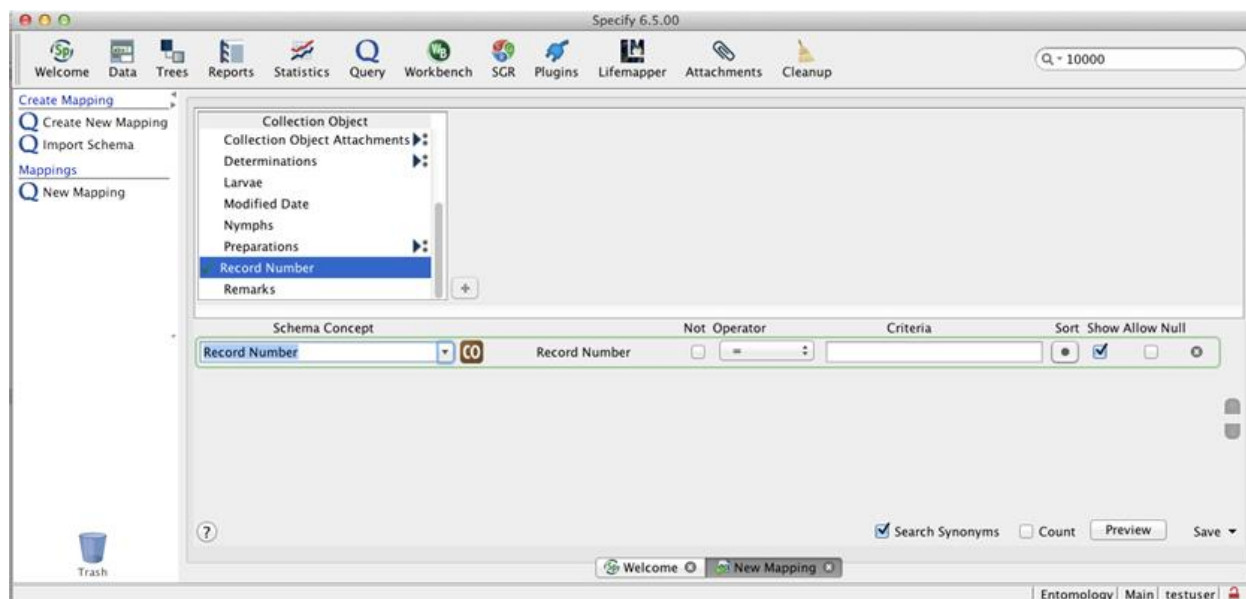
Figure 1. Creating a new Export Data Schema in the Specify Schema Mapper. Fields are chosen in the top window from the Specify database are added as Schema Concepts to the draft Export Data Schema below. Selection filters and other parameters can be set for each concept and when complete the Export Data Schema is saved and will appear under the "Mappings" list in the left window pane.

Delete a term in the Export Data Map by clicking the ⊗ *(remove field)* button. Query Operators and Expressions can be used to narrow the results. These function the same as the in the Query Builder, except here they limit the query results to data to be exported. The 'Show' check box determines whether that field and its data are exported or not. (Sometimes a field is used to filter the set of records to be exported but the field itself is not used in the export file, see below). The 'Allow Null' check box enables records to be included in the export if the contents of that field are 'null' – even when the fact that they are 'null' would not meet the specified filter condition. (Occasionally, one wants to export records which contain no value in a field which would be otherwise filtered by some criterion.)

3.2 Define a New Export Data Map using a Predefined Data Schema

This Export Data Map creation workflow assumes that you are using an existing export schema obtained from a database aggregator or other project which specifies the fields and format they require for your data. Producing a DarwinCore Archive file is an example.

Log into Specify and open the Schema Mapper within the Specify application by choosing: *System > Schema Mapper*.

  ☝ **You must use a Specify user account at the Specify administrator-level to be able to run the Schema Mapper.**

First, import the predefined schema. You import a schema by clicking the Import schema button and importing a schema file. Then choose a predefined schema from the *Create New Mapping* dialog list. The Schema Mapper editor window will open with a list of Specify fields in the top section and a list of data types, referred to as Schema Concepts in the bottom section. In figure 2, a predefined schema was chosen, and the data concepts it contained are shown in the 'Schema Concept' column. Scrolling through the lower window pane will show more concepts, and clicking on the small disclosure arrow in the Schema Concept column will permit choosing another of the predefined concepts from that schema to some other for the selected Specify field. Only unmapped schema concepts are available in the list.

Specify attempts to automatically map database fields to export schema concepts based on matches between field and concept names. Some Schema Concepts may already be mapped by Specify after choosing an export schema, but they can be edited and changed if desired.

Here is a list of existing data schemas which may be relevant to your collection if you participate in community data portals which aggregate data from multiple collections:

- Ocean Biogeographic Information System (OBIS): http://www.iobis.org/data/schema-and-metadata, schema document: http://www.iobis.org/obis/obis.xsd

- The Paleontology Portal, schema document: http://data.paleoportal.org/schema/conceptual/1.0/paleoportal.xsd

- Access to Biological Collections Data (ABCD): http://wiki.tdwg.org/ABCD/, schema document: http://rs.tdwg.org/abcd/2.06/ABCD_2.06.xsd

- Taxonomic Databases Working Group (TDWG) Darwin Core (DwC) Standards: http://rs.tdwg.org/dwc/. The DwC schema is included with Specify programs and is located in the Specify installation directory (../specify/config/). DwC is used by the Global Biodiversity Information Facility for formatting the contents of Darwin Core Archives files which are the basis for GBIF uploads. (See: http://rs.tdwg.org/dwc/ for Darwin Core standard information).

External data schema in ".xsd' file formats can be downloaded and added to the Specify /*config* directory and then used in the Schema Mapper by clicking on *Import Schema*.
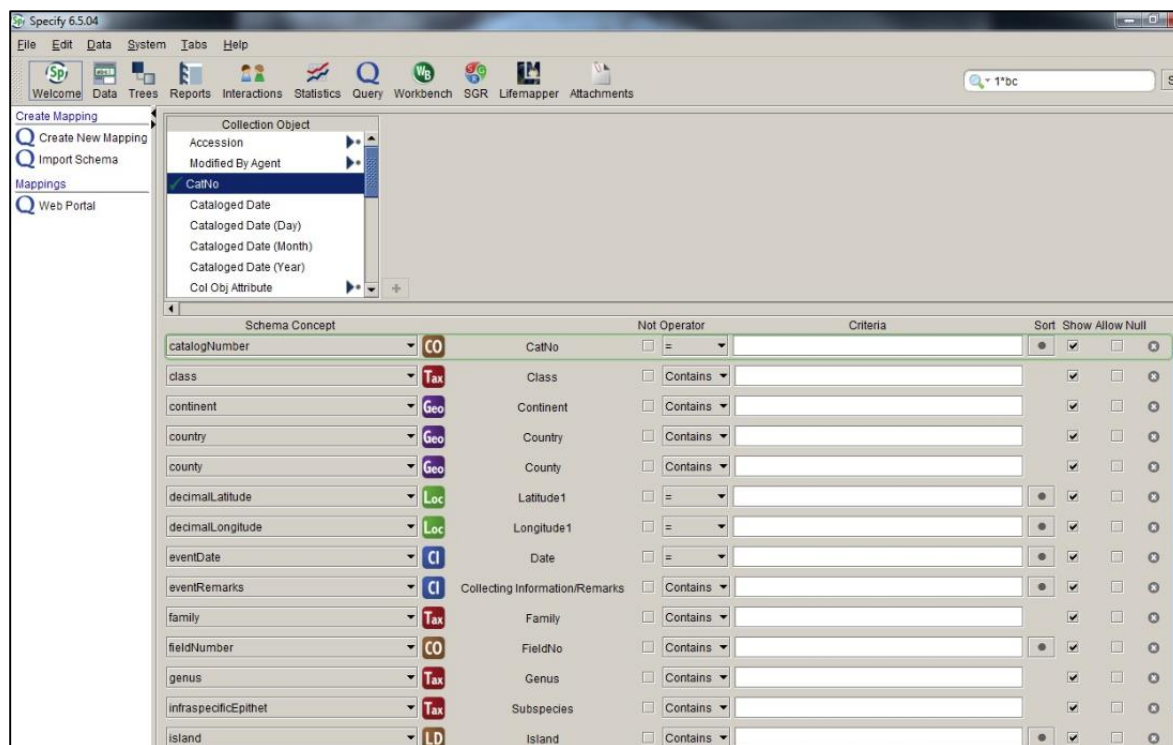
Figure 2. Specify Schema Mapper showing the use of an export map for mapping Specify data fields to create an Export Data Schema. The Schema Concepts were preloaded from the DarwinCore file format standard definitions, the second column shows the Specify table icons and field names which are mapped to those concepts.

Add a field by selecting one from the Specify database and then clicking the ✚ (add) button. The Schema Concept will initially be defined as 'Not Mapped'. Choose an entry from the Schema Concept drop down list to map the data field to that concept, or leave it as 'Not Mapped' if the concept is not in the list of Schema Concepts (for example, for concepts that are used purely as filters for exporting data). *Note:* it is not necessary to include a collection object attachment field in Export Data Map, references (links) to collection object attachments are automatically included in the exported files if the attachment server is used for attachments.

Delete a term in the export schema by clicking the ⊗ *(remove field)* button. Query Operators and Expressions can be used to narrow the results. These function the same as the in the Query Builder, except here they limit the data to be exported.

3.3 Adding Data Fields to an Export Data Map to Filter Exported Records

The content of a GBIF Darwin Core Archive file (https://code.google.com/p/gbif-ecat/wiki/DwCArchive) of uploaded collection data and content of data sets uploaded to the Specify Web Portal must contain unique collection object data records--duplicate records are not allowed. In the case of Specify, for the Web Portal to successfully ingest records from a Specify database, the unique ID which identifies the collection object record, typically the catalog number, may not be duplicated among the export file record identifiers.

Because of the presence of one-to-many data table relationships in Specify, queries would normally return multiple copies of a flat specimen record if, for example, a specimen record was associated with the names of two or more collectors (one query result row for each collector name), unless the contents of data tables related one-to-many to Collection Object are formatted within Specify to be aggregated into a single field.

One can add a formatted table from the Specify database structure display on the top of the screen, by double clicking on the table name to take advantage of predefined formatted and aggregated values for that table. For example, selecting the Collectors table by double clicking 'Collectors' will concatenate the names of all of the collectors for a specimen in a single field. Many tables have default aggregators pre-defined to concatenate values in multiple records linked to a collection object record in a one-to-many relationship. To find out which tables have aggregators already pre-defined, log in with a Specify admin user account and open the Schema Configuration tool from the System menu. Those tables with aggregators defined will be identified by having a name in the 'Table Aggregation' field in the Schema Mapper form. If no aggregator is defined for a one-to-many related table, information on how to create table aggregators is available in the online help within Specify, or contact the Specify Project for assistance.

You may prefer to export only one of the values from a one-to-many data relationship instead of aggregating all of the values. For example, for taxon name determinations, you may wish to include only the current determination, so that one and only one taxon name record is exported with a collection object record in the output file. One can do this by adding the Specify data field "Current" from the Determination table and setting it to "*Yes or Empty" in the Export Data Map.* Doing that will eliminate multiple determinations (and taxonomic names) and multiples of the same collection object record from being exported.

In some cases it is advantageous to add other fields the Export Data Map in order to filter out certain records from the output file. For example, an environmental status field might be added to filter out specimens of species which are threatened or endangered. By unchecking the "*Show"* checkbox on the export schema concept, the query results will not include that filter field and its contents in the resulting data. The 'Current' field in the Determination table would be a good example of a filter field which would produce the taxon names considered current but the field 'Current' itself would not be exported or show in the web portal display.

3.4 Check the Schema Mapping

When and Export Data Map seems complete, click the *Preview* button in the lower right corner of the Schema Mapper to display a 120 record sample of the specimen records resulting from your mapping. If your data contains duplicate records, a message will remind you to correct this problem (see discussion of removing duplicate records above or contact the Specify Project for assistance). If all the concept mappings are correct, save the Export Data Map by using the *Save* button give the EDM a file name. It will appear in the left window pane entitled "Mappings". You must close the Export Data Map before moving to the next step.

4. Create an Internal Cache (Copy) of your Export Data

When your Specify data fields have been mapped successfully to an Export Data Schema, the next step is to create the internal data cache. That cache or copy of your data which is formatted and filtered by the EDM, is created using the Specify Data Exporter a software program located in the Specify installation directory (../Specify/bin/DataExporter).

Important Note on Required Account Credentials for the Specify Data Exporter

- ☙ Running the Data Exporter requires a Specify user account at the administrator level.
- ☙ The Specify 'IT User' credentials are required when using an Export Data Map for the first time to create the internal cached copy, or whenever an Export Data Map is changed. The Specify 'IT User' account was set up when the Specify database was first created, the account has MySQL privileges to create new data tables. For normal updates to the internal cache (updating only the data and not the Export Data Map used), the Specify 'IT User' account is not required.

1. Double-click the Specify Data Exporter (DataExporter.exe) to launch it and log in using an administrator-level Specify account username and password. The Data Exporter window opens as a three-column table with action buttons along the bottom.
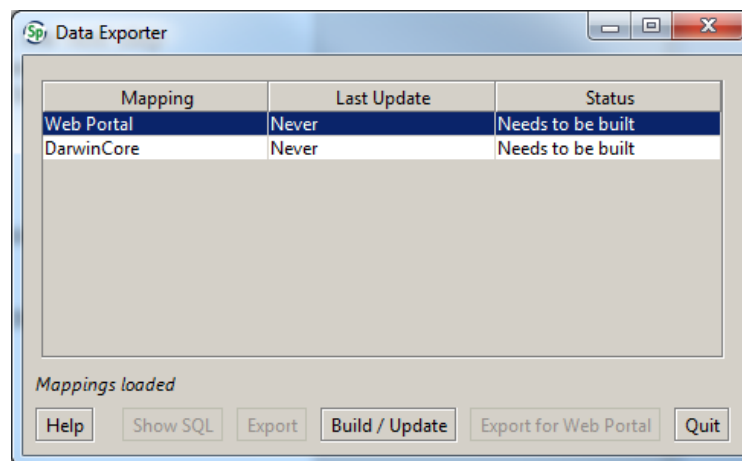

Figure 3. The Specify Data Exporter window shows two existing mappings or Export Data Maps, and their status having never been updated and needing a new build. Using the Build/Update button would build a new cache or update an existing one with database records added or modified since the last update, for the Data Map selected.

The <u>Mapping</u> column of the table shows the name of the Export Data Maps you have created.

The <u>Last Update</u> column shows the date and time the internal export cache was updated.

The <u>Status</u> column indicates whether the cache needs to be built for the first time, is 'Up to date', or is 'Out of date'. The internal cache is out of date if any data records within Specify have been deleted, edited, or added to the database since the last cache update. This column also shows the number of records which are out of date. The number reflects the database contents at the time the Data Exporter program was launched; it will not reflect changes to the database that occur while the Data Exporter is running.

2. Build the Internal Cache. If the cache for an Export Data Map requires an update, select it and then click the Build/Update button. You will be prompted for the login credentials of the Specify 'IT User" account only if the Export Data Map is new or has been changed since the last build.

For internal cache tables which have already been built, only out-of-date rows of data are updated. This is significantly faster than an initial cache build which can take a few hours for a very large database using an average workstation. Updates to internal caches are not allowed if Export Data Maps are open. Close any Export Data Maps in the Schema Mapper before attempting an update the internal cache with the Data Exporter program.

3. Click *Export for Web Portal* and choose or create a directory for the resulting .zip file to be saved. The .zip file is a package of files and a directory which contains your output data and files containing configuration information for the Web Portal.

   💣 When naming an export .zip file use only alphanumeric characters and do not include spaces.

5. <u>Load Data into the Web Portal</u>

The content of a Specify Web Portal is a snapshot of the information in the database at the time an export cache file was built and exported. Any changes made in a Specify database after an internal cache build and a .zip file export Web Portal export will not be present in the internal, flat cache copy of your data nor in the exported data package until the internal cache is updated with the Data Exporter program.

Once an export data .zip package has been produced using the Data Exporter (or command line tool – see below) the Web Portal itself needs to be updated to make the new data accessible through the Portal's web pages.

The Web Portal is updated with these steps:

 1. Copy the exported .zip file to the "/specify_exports" directory on the Portal server.

 2. Re-build the SOLR application by issuing the following command while working in the server directory containing the Makefile for the web portal installation:

```
make clean && make
```

 3. Re-install the web portal by typing the following command:

```
sudo make install && sudo invoke-rc.d Tomcat 7 restart
```

It is possible to automate the Web Portal updating process on the server by making changes within Tomcat 7. Refer to the README file located on the Specify Web Portal GitHub site: https://github.com/specify/webportal-installer.

5.1 Update and Export the Cache using the Data Export Command Line program

The Specify Software Project distributes another software utility, the Data Export Command Line tool (ExpCmdLine.exe) to automate updating data in a Specify Web Portal.

Once an internal cache (copy) exists and the Data Export Map is current (has not changed since the last build) the cache can be updated and exported to a Web Portal at the command line by using the Export Command-Line tool (ExpCmdLine.exe) found in the ../Specify/bin/ directory. Using the command-line tool offers the ability to write scripted automatic updates to the Portal.

> ☀ If an Export Data Map has changed the cache must be re-built using the Data Exporter program. If a rebuild of an internal cache copy is necessary the Export Command-Line program will not rebuild nor update the cache.

See the separate documentation on the Specify web site entitled "Using the Specify Command Line Data Export Program (ExpCmdLine)" on the specify Software Project web site, for how to script updates to a Specify Web Portal with data from a Specify database.

6. Using the Specify Web Portal

Browsing data in the Specify web portal begins with a data search. Search results can then be examined in three different views.

6.1 Search

Express Search functions much like it does within the Specify application and is most useful quick search on single terms, strings or numbers. For express searches on phrases containing spaces, use quotes, e.g. when looking for collection objects from Costa Rica, search on: "Costa Rica". The * character can be used as a wildcard for stemming searches both before and after a search term and within quoted strings. Searching on "*Rio Sucio*" will find the string "Rio Sucio" anywhere it might be embedded within a locality description with words before or after it.

Advanced search allows users to query the fields included in the mapping file. These fields can be hidden in the search but still displayed in the results by editing the web portal settings.json file.
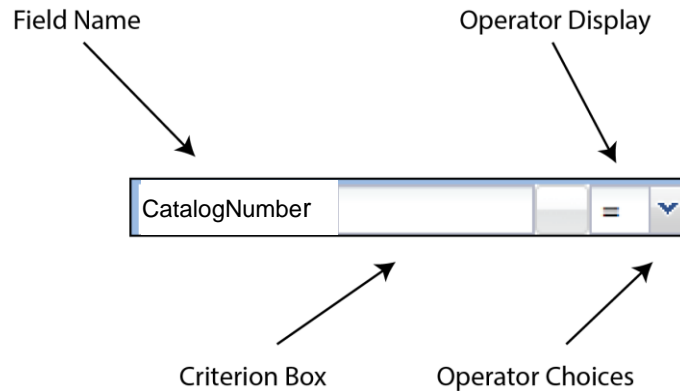
Figure 4. The anatomy of a search field display from the left side Search Window pane in the Web Portal.

Table 1. Available query operators for Specify Web Portal by data field type.

Operator	Result	Field Type
=	Finds all records for which field values are equal to the specified criterion.	Number, Date, Text
>	Finds all records for which field values are greater than the specified criterion.	Number, Date
<	Finds all records for which field values are less than the specified criterion.	Number, Date
> =	Finds all records for which field values are greater than or equal to the specified criterion.	Number, Date
< =	Finds all records for which field values are less than or equal to the specified criterion.	Number, Date
Between	Finds all records for which field values are between the specified values.	Number, Date
In	Finds all records for which field values are included in the list of criteria. (example criteria; Northern, North, No. or 32, 35, 36, 39)	Number, Text
Like	Allows a wildcard character (*). Finds all records for which field values matches the specified pattern expression. (example; *ist)	Text
Contains	Finds all records for which field values contain a specified character string.	Text
Empty	Find all records for which the field contains no data. A Condition that is set to Empty will filter out records that contain any value in the field and return only records which have nothing in that field.	Number, Check Box, Text
TRUE	Finds all records for which field value is True.	Check Box (Boolean)
FALSE	Finds all records for which field value is False.	Check Box (Boolean)

6.2 Search Tools

Search tools are located in the side bar on the left side of the Web Portal window.

Geo Coords limits search results to records which have data in the latitude and longitude fields.

Images limits search results to records which have an attached image.

Fit to Map is an option available in the Map View. Use Fit to Map to execute a spatial query for georeferenced specimens localities contained within the visible portion of the map. Check this option, then increase or reduce the size of the map window, adjust the position of the underlying map image to delimit the spatial area to be searched, and press the 'Search' button to obtain just those specimens from that spatial extent.

6.3 Data Views

Data can be viewed as records, images, or points on a map by selecting the appropriate tab at the top of the browser window. Double clicking a row in the records tab, an image in the image tab, or a map point in the map tab will pop open a floating specimen detail data window.

The **Records** tab (Figure 5.) displays records in a grid view of columns and rows.

Figure 5. Specify 6 Web Portal Detail Display

Clicking on the small disclosure triangle toward the right of any column header entry, which appears when you hover over the column header, will pop open this menu:
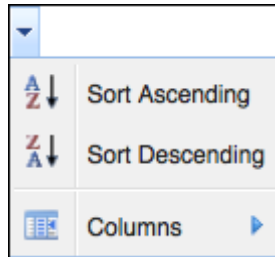

Figure 6. Column Header Pop-up Menu

Each data column can be sorted in ascending or descending order, and by moving the mouse over the *Columns* entry at the bottom of the pop-up, all fields can be selected or hidden from view. Field columns can be re-ordered left-to-right by dragging the column header to a new location.

The **Images** tab (Figure 7.) displays only the images associated with the records resulting from the current search. If an image/attachment server is not accessible to the web portal this view is not available.
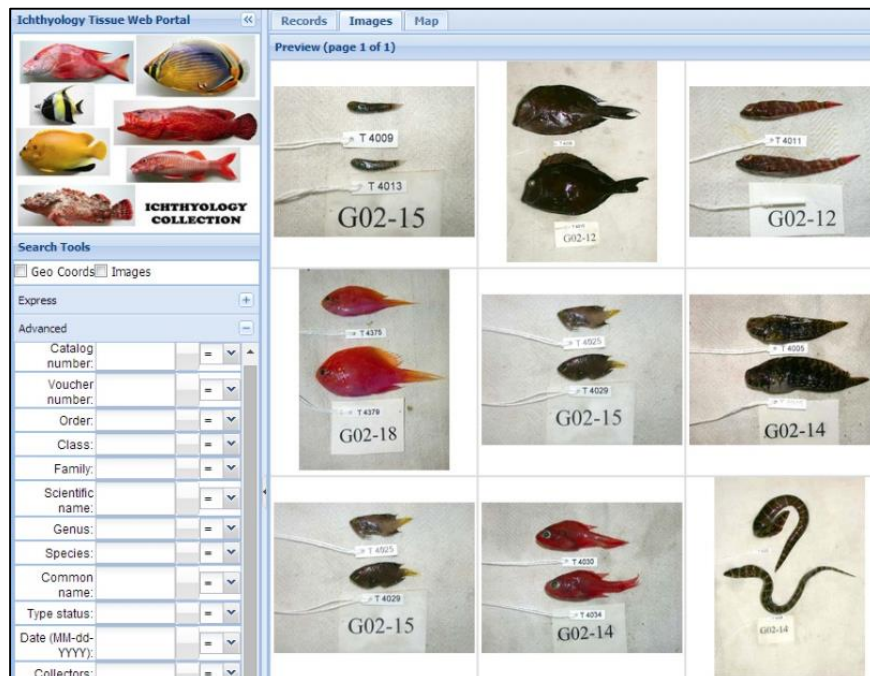

Figure 7. Specify 6 Web Portal Image Tab Query Results

The **Map** tab (Figure 8.) plots the location by latitude and longitude of the georeferenced specimens described in the query results. The Fit to Map option executes a spatial query for collection objects bounded by the extent of a map visible in the window.
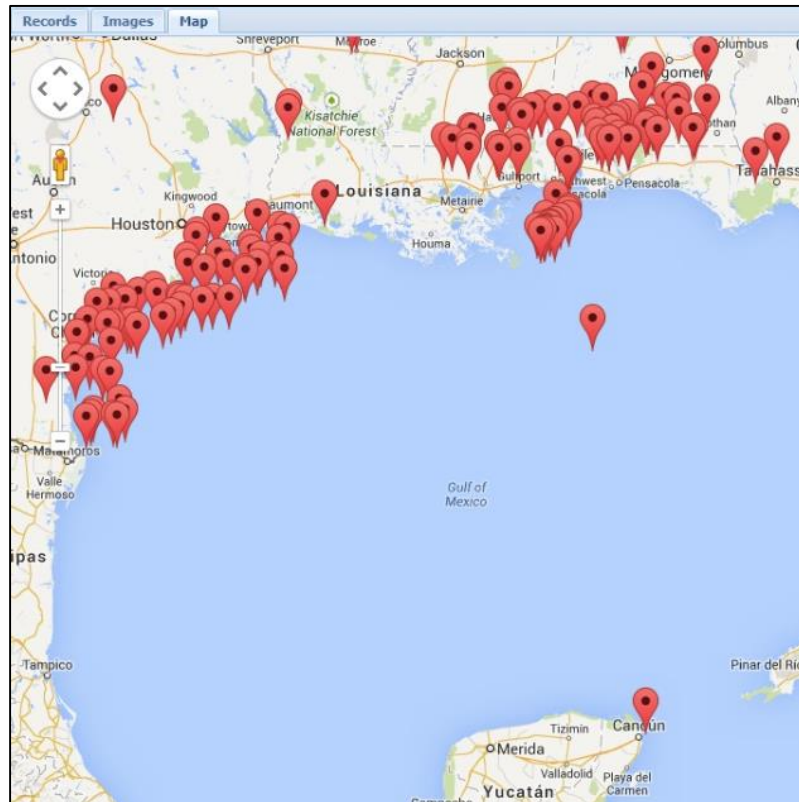

Figure 8. Specify Web Portal Map Display