Columbia Basin Water Hub

User Manual

Site development by Genki Maps

Table of Contents

Credits

This user manual was adapted from the Help Pages for the Skeena Salmon Data Centre CKAN site, operated by the Skeena Knowledge Trust (SKT), with their explicit permission. The original help pages for the SKT CKAN site are available at: https://www.manula.com/manuals/skeena-knowledge-trust/skeena-knowledge-trust/1/en/topic/skeena-salmon-data-centre. Full credit goes to the Skeena Knowledge Trust (SKT) for the original content of their Help Pages.

Please note that this user manual is intended as a guide to the Columbia Basin Water Hub CKAN site, and in particular the customized features of this site. It is a supplement to, not a replacement for, the general-purpose CKAN documentation, which can be found at: https://docs.ckan.org/en/2.8/user-guide.html.

1 Introduction

The Columbia Basin Water Hub (CBWH) is operated by Living Lakes Canada, an organization working towards the long-term protection of Canada's lakes, rivers, wetlands and watersheds. CBWH is a web-based data portal platform, running a customized software system called CKAN (https://ckan.org/). The Comprehensive Knowledge Archive Network (CKAN) is a web-based application for the storage and distribution of open data. It is a powerful data catalogue system, used by numerous governments, community groups, and institutions with a need to share data.

The CBWH website, at https://data.cbwaterhub.ca/, was developed and customized by Genki Maps, a geospatial and data management software consultancy. Customized features of the CBWH include auto-loading of data from Excel tables, a map search feature, and a metadata schema specific to the water management needs of Living Lakes.

1 Core CKAN Concepts

This section describes the way in which data and users are organized and displayed in CKAN. The core concepts or "structures" used to organize data are Datasets, Resources, Organizations, and Groups. We also discuss Metadata, which is additional information about each dataset and resource. These are discussed in the following sections.

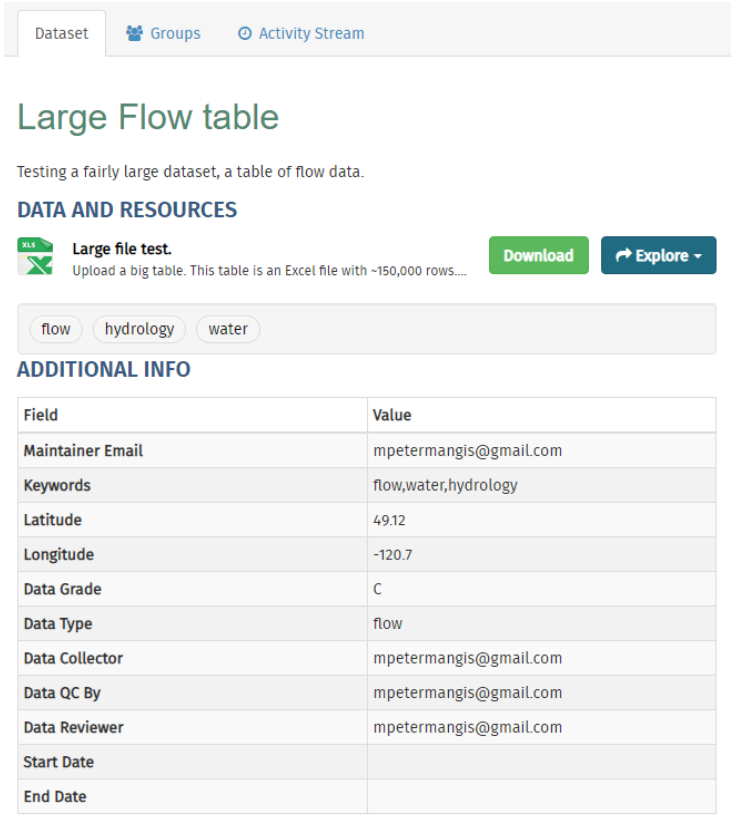
1 Datasets and Resources

In CKAN, a *dataset* is a collection of related data and metadata. Each dataset may contain one or more *resources*, which are either files uploaded to CKAN, or links to data stored elsewhere. A resource could be a PDF or Word report, a map, a data table, or something else. Both datasets and resources have their own metadata. For the CBWH site, the majority of resources will be Excel files

containing tables of various kinds of water monitoring data. Typically, one monitoring station would correspond to one dataset in CKAN. All data tables and reports pertaining to that station would be added as resources to its dataset.

A dataset contains two things:

1. Information or "metadata" about the data. For example, the title and publisher, date, what formats it is available in, what license it is released under, and so on. CBWH uses a custom metadata schema, with metadata fields specific to water monitoring, in addition to the standard CKAN metadata fields.
2. A number of "resources", which hold the data itself. Resources also have their own resource-level metadata, which is independent from the dataset-level metadata.



Default view of a dataset and its resources in CKAN.

1 Organizations

Each dataset is owned by an organization, which is a collection of users who belong to a common group. This could be users from a First Nation, users from a specific government department, and so on. The CBWH is expected to eventually have dozens of organizations. Each organization can have its

own workflow and authorizations, allowing it to manage its own publishing process. A large organization can break up its data by department, allowing each department to be a separate organization within the CBWH.

Each organization can have a user assigned as an administrator. An organization's administrator can add individual users to it, with different roles depending on the level of authorization needed. Each dataset in CKAN belongs to only one organization.

A user in an organization with Editor permissions can create a dataset owned by that organization. By default, this dataset is initially private, and visible only to other users in the same organization. Only logged-in users who are members of the dataset's organization can see private datasets. When it is ready for sharing outside the organization, it can be made public at the click of a button. A Public dataset is visible to all users of the site.

1 Organization Permission Levels

Users added to an organization can have one of three roles: Admin, Editor, or Member.

Member: can view the organization's private datasets, and all public datasets.

Editor: all of the above, and: create and delete datasets within the organization.

Admin: all of the above, and: add and delete users, change user roles, change metadata about the organization.

2 Groups

Groups in CKAN does *not* refer to user groups. A "group" in CKAN is really more of a data theme—it is typically used to group thematically-related datasets. Whenever you see Groups, think of Themes or Categories. Groups are meant for the community of site users to collect related datasets together into themes such as Water Quality.

The Groups used by the CBWH will reflect the type of waterbody that a dataset is related to. The current groups are Climate, Groundwater, Lakes, Snow, Streams, Wetlands, and Glaciers. These groups are intended to have a 1:1 correspondence to the Data Type field in the custom metadata, although this relationship is not set in stone.

1 Group Permission Levels

Editor: add or remove datasets from the group.

Note that a dataset can belong to more than one group, unlike organizations.

3 Metadata

In CKAN, datasets have a default *metadata schema,* which is a set of fields with information describing the dataset. These include dataset title, summary, maintainer email, parent organization, and keywords, among others. The CBWH has been customized with additional metadata fields related to water monitoring data: Latitude, Longitude, Data Grade, Data Type, Data Collector, Start and End dates, and others. The custom metadata schema is currently fixed, but it can be changed as needed. The list of keywords is restricted to a controlled vocabulary. The Data Type field is categorical, and has a set of data types determined by Living Lakes to best meet the needs of the organizations that will be uploading data. The list of data types corresponds to the Groups in CKAN and also the available layers in the Map Search interface.

2 Publishing Data

This section describes how to publish your data on the CBWH. Note that you can only publish data if your account has Editor or higher permissions level.

4 Preparing your data

The CBWH includes a custom extension that automatically loads data from uploaded Excel and csv files, with field type detection, to an internal database within CKAN, called the Datastore. The Datastore allows users to preview the contents of tabular data, extract a subset of that data using a query, and download data in other structured formats such as csv or JSON. For more information on the CKAN Datastore, see: https://docs.ckan.org/en/2.8/maintaining/datastore.html.

1 Structuring Tabular Data

The Datastore loading extension handles data in Excel and csv format, and is designed to be as flexible as possible. There is no requirement for column names or types to match any pre-existing schema for water management data. However, there are a number of restrictions that apply to the format and structure of tabular data. It is impossible for the plugin to bypass these restrictions, so any table that does not match the following list of restrictions will not be loaded to the CKAN Datastore. If this happens, the uploaded file will still be published in CKAN and be downloadable by authorized users, but the Datastore feature specifically will not be enabled for that table. Please review the following restrictions prior to publishing your data.

- The excel or csv file must have a contiguous rectangular block of data (rows and columns) with a header row. There must be only one header row. There must not be a gap between the header row and the first row of data, even if this improves human readability. Multiple header rows embedded in the data table will cause all columns to be interpreted as text columns.

Multiple blocks of data separated by gaps of blank cells will likely cause data loading to fail. Before loading an Excel file, ensure that all hidden columns and rows are visible, and check for any issues with the structure of the table.

- Each column must contain only a single data type (such as integer, decimal number, text). The data types must be simple, and not composite. For example, two columns Latitude and Longitude, each containing decimal numbers is OK. By contrast, a single column containing a latitude-longitude *coordinate* will not be interpreted correctly. Columns may contain null values, with the following caveats: if using Excel, a null value is represented as an empty cell with no data at all, not a space character, zero, underscore, a formula that evaluates to a blank cell such as =””, or any other data. If using csv, a null value is represented as a pair of commas with no space or quote marks between them: ,, Any other representation of null in a csv is likely to be interpreted as a text object, and cause the entire column to become a text field.
- Excel files must not have any graphs, pictures, comments, formulas, or other objects embedded in the data table. Avoid using merged cells anywhere in the file.
- The header row does not have to be on the first row, but there must be a header row somewhere in the file. You will have to provide the Excel/csv row number of the header row when uploading your file; it cannot be reliably auto-detected, and cases where the header row is not on the first row are sufficiently common that a default of assuming that the first row is the header row is not sensible. The header row number starts from 1 on the first row; it is not zero-indexed. If you leave the header row field blank, or enter an invalid row number, the file will be published as a resource in CKAN, but it will *not* be loaded to the CKAN Datastore.
- The header row must contain only SQL-compatible column names, meaning only alphanumeric characters and underscores (no spaces, brackets, commas, hyphens, or other special characters), and less than 63 characters long. Internally, all column names will be converted to lowercase before loading the table to the Datastore. Note that this only applies to the copy of the table that is loaded to the Datastore; the original column names in your csv or Excel file will remain as they were when you uploaded the file.
- In CKAN, a resource can only have one table loaded to the Datastore. If the Excel file has multiple spreadsheets, only the first one will be imported in the Datastore. However, the original Excel file will not be modified; if you have multiple sheets in the uploaded file, they will remain unchanged when users download the original file. If you have an Excel file with multiple sheets that you want to have loaded in the Datastore, please copy each sheet to a new Excel file, and create a new resource for each of these files.

2 Creating a Dataset

On the CBWH site, on the Dataset page, click Add Dataset. If you do not see the Add Dataset button, your account does not have the access level required to create datasets.

This is a two-step process. In the first step, you will see a series of metadata fields, including basic fields like Title and Summary, along with others specific to water monitoring. If you are creating a dataset for which some of these fields do not apply, simply leave them blank. The metadata schema is designed to be inclusive of all possible data types that might be published by Living Lakes, so it is likely that some of the fields will be blank for every dataset you create.

1 Upload and Wait for Publication

On the next page, either click Upload to upload a data file, or click Link to add a link to a dataset that is hosted on an external website. If uploading a tabular data file, please be sure to enter the header row number. Also be sure to correctly enter the format. The format can be detected from the file extension, but this is not always reliable.

If you upload an Excel or csv file with a header row defined, the Datastore loading extension will be triggered in the background, and will attempt to load your data to the Datastore. While this is happening, the site will redirect you to the resource page for the file you just uploaded. It is likely to take anywhere from a few seconds to many minutes before your data becomes available in the Datastore. One of the metadata fields in the resource page shows the status of data loading. This field will contain one of the following: an in-progress message, a completed message with the number of rows loaded, or an error message. Once the process is complete, refresh the page to see your table in the datastore, which shows roughly the first 25 rows in a table embedded at the top of the resource page in CKAN. If instead of the data table preview, you see the message “There are no views created for this resource yet” it means that the Datastore load failed for this table. Scroll down to the Loading Status section below and there may be an error or status message. If this still does not clarify the reason that data loading failed, please contact a site administrator.

1 File Size Limit

There is an upload size limit of approximately 900MB per file. If you need to upload files larger than this, please contact a site administrator to arrange an alternative data workflow. Please remember to be mindful with zip files that may cross the 900MB size, and contact us with any questions. Generally, data in Excel will consume less space than the same data in csv format.

2 Retrieving Data

We expect that there will eventually be a large number of datasets in the CBWH, which could make finding relevant data challenging. This section describes ways to carry out data searches, narrow results by filtering, download data, and query subsets of data.

1 Searching for data

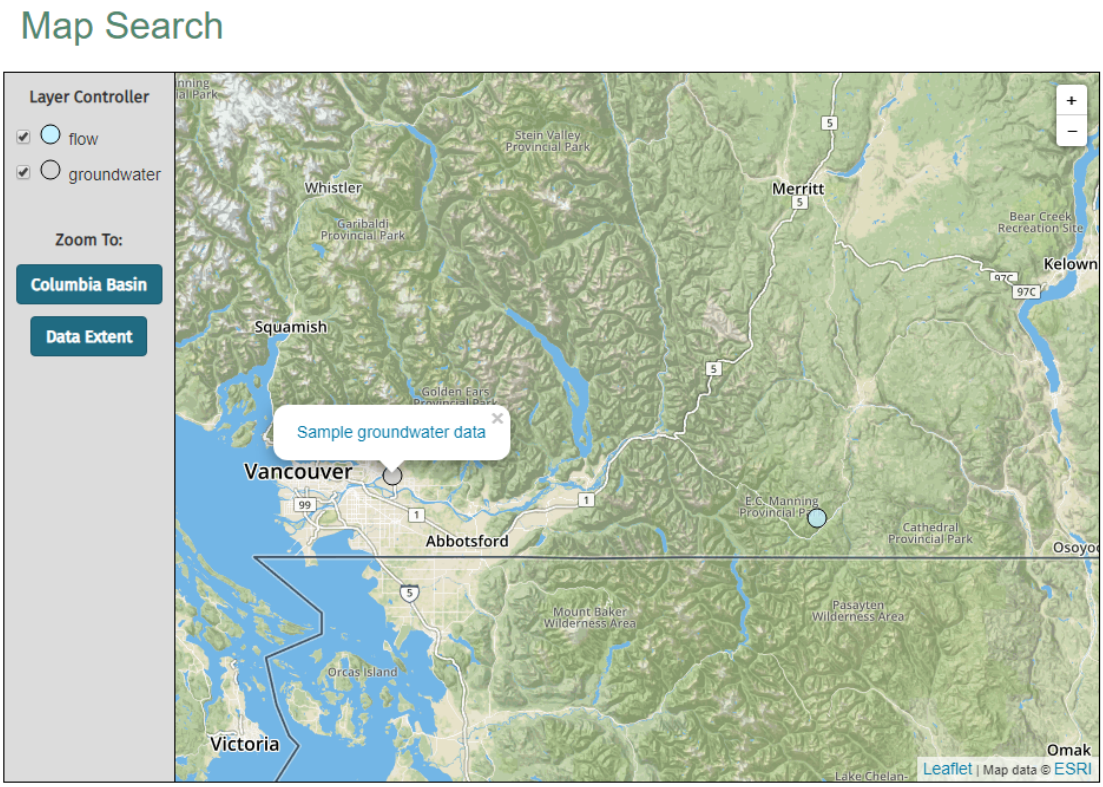
There are several ways to search for data in the CBWH.

1 Search by Map

This is an entirely new feature, customized for the CBWH. Click the Map Search link in the menu bar. This will bring up a map display with a legend. The points on the map correspond to all datasets that have a latitude, longitude, and data type in their metadata. One point is shown for each dataset, but a single dataset may contain multiple resources. For example, a dataset for a water quality monitoring station at a fixed location might have a data table (resource) for each year of data collection. The points that you see on the map are limited to those datasets that your user account can view. For example, a private dataset in an organization that you are not a member of will not appear on the map.

As with most web maps, use the mouse to pan around and the mouse wheel to zoom in and out; there is also a zoom control at top right. The map by default is zoomed to the Columbia River basin. There are buttons on the left to zoom to this extent, or to the full extent of all datasets. The Layer Controller on the left shows the list of all data types in the datasets that you can view. Use the check mark next to each data type to turn the layer on or off. If a dataset has more than one data type, there will be a stack of points. Use the layer controller to filter only the data types that you would like to see.

Clicking on a point on the map brings up a small pop-up with the dataset's title and a link to the map. Clicking the link in the popup takes you to the dataset page. If more than one dataset has the same latitude-longitude coordinate, the popup will contain links to all datasets at that coordinate.



The Map Search view on the CBWH site.

2 Browse Datasets

Clicking the Datasets menu item takes you to the standard browser view of datasets. You will see a list of datasets with a short description of each. By default, these are sorted by Relevance, but if you have not entered any search keywords, the default arrangement is the same as Last Modified. The Order By menu at the top-right has more options for sorting this list of datasets. Datasets can be filtered on the left side of the page by Organizations, Groups, Tags, Formats or Licenses. Similar to the map search, you will only see those datasets that your user account has permission to view.

2 Text Search

The text box at the top of this view has a fairly good full-text search engine. The indexing is based on significance of keywords in the document. So, for example, if you search for "water quality," results with that phrase in the title or near the beginning of the summary will appear highest in the list.

The following search operations may be used to perform more precise searches:

Search Operator	Example	Effect
Minus (-)	-columbia -estuary	Excludes terms or phrases from the search.

	water quality	In this example, the search returns datasets for the search *water quality* without the terms *columbia* or *estuary* in the returned results, which helps to exclude unwanted information.
Quotation Marks (“”)	“water quality”	Makes an exact-match search of the quoted phrase. In this example, the search returns datasets that contain the whole quoted phrase *water quality*, which helps refine search results.
AND	Canal flats AND water quality	Finds datasets related to multiple phrases. In this example, the search returns datasets that are related to both *Canal flats* and *water quality*, which helps narrow the search.

It is possible to combine the search operators to create more customized searches.

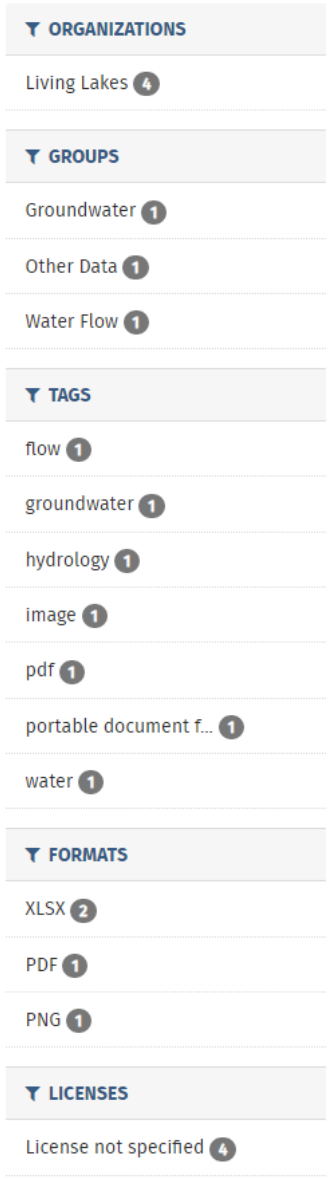
1 Browse by Group (aka Theme/Category)

Categories are broad topics of interest used to group similar datasets and include Climate, Groundwater, Lakes, Snow, Streams, Wetlands, and Glacier. Some datasets are associated with more than one category. Categories can be found at the bottom of the home page or under the Groups menu.

2 Browse by Organization

Click the Organization menu, then click on the organization of choice to see all *public* datasets belonging to the organization. If you are a member of the organization, you will also see its private datasets.

3 Filtering search results



After searching or browsing for datasets, you may wish to filter the list to further narrow the results. A search can be filtered by resource file format, group, organization, tag, or license, or any combination of these. Filters are found on the left side of the search page.

1 Filtering by Format

Since resources can be provided in multiple formats, clicking on one format in the left panel will filter the results to show only datasets with resources distributed in that format. You may further filter the results by selecting additional formats; only datasets containing resources with both file formats selected will be returned.

2 Filtering by Group (Category)

Since datasets may be related to multiple collections/groups, clicking on one group will filter the results to show only datasets of that collection/category. However you may filter further to show those resources from the originally selected group that also fall into and one or more additional groups.

3 Filtering by Organization

Each dataset is linked to a single organization. While you can filter by organization, it is not possible to select a second organization filter, since datasets cannot belong to more than one organization.

4 Filtering by Tag (Keyword)

Each dataset can have multiple keyword tags associated with it. You can filter search results by tags; as many tags as desired can be used as filters.

5 Filtering by License

Datasets in the SSDC may have a license associated with them. Clicking on one of the license types will filter the results to only show datasets with that license. Datasets without a license type are tagged as “License Not Specified.”

6 Combining Filters

Multiple categories of filter can be chosen. For example, a search can be filtered by tags, file formats, and groups. The selected filters will be displayed at the top of the search page. The filters are *inclusive*, meaning that only datasets matching all selected filters will be shown.

To clear selected filters, click the X next to the filters you wish to remove, either at the top of the search results page, or in the filters panel on the left side of the page.

7 Other Filtering Tips

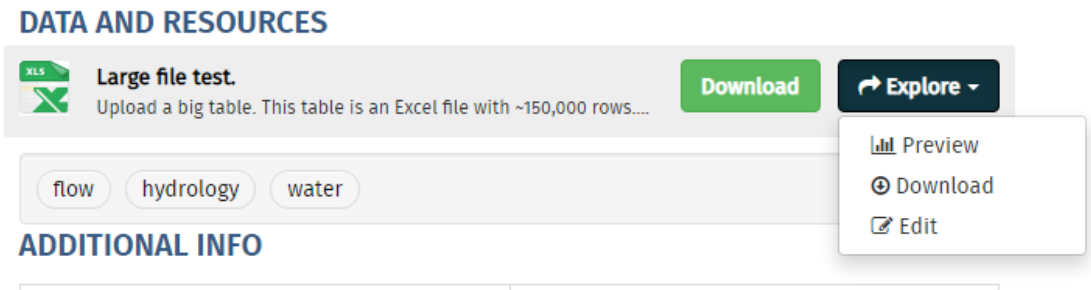
If you want to use a filter that is not listed, click on the "Show More …" button at the bottom of each filter category. If the desired filter term still does not appear, add the topic of interest into the search bar. The filters most related to the search topic should now appear in the list of filters.

4 Downloading data

Once you have found a resource of interest, you can download it from the CBWH.

1 Download From the Dataset Page

Every dataset has an ID, which is a URL-friendly version of the name in lowercase. You will see the dataset ID at the end of its url, for example a URL ending with /dataset/flow-data is the "dataset page" for the dataset "Flow Data." The easiest way to download data is by clicking the green Download button for data of interest. This allows you to download the data in its original format (usually Excel, for the CBWH). On the Explore drop-down menu, there are additional options. The Download option on this menu is the same action as the green Download button. Preview brings you to a page where you can preview the data table (if the data is tabular) or file (preview is supported for PDF and common image formats).

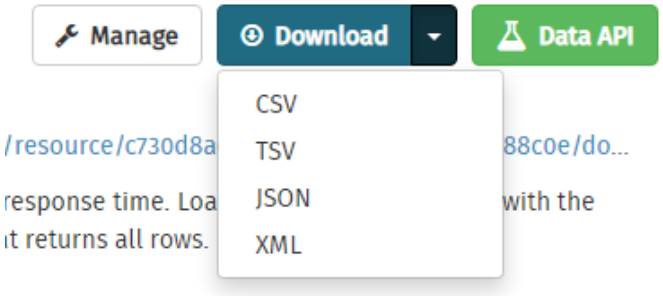


Download options from the main dataset page.

2 Download From the Resource Page

To open the Resource page, click the link for the name of the data file, in this example "Large file test." This brings you to the resource page for a data file. If the data is tabular, and has been successfully

loaded to the CKAN Datastore, you will see a preview of the first few rows of the table. If the data is a PDF or image file, there will be a preview below. You can also download data from this page. To do so, simply click the Download button near the top-right, avoiding the down arrow at the right edge of this button. If the data is tabular, you also have additional download options in different format, which can be accessed by clicking the down arrow at the right side of the button. However, downloading data in its original format is recommended.



Download options from a resource page within a dataset.

3 Querying Data

For tabular data, it is possible to query a subset of the data, instead of downloading the entire file at once. This can be useful, for example, if you want only data collected within a certain date range, or only values that exceed some threshold. In the CBWH, supported Excel files are automatically loaded into the CKAN Datastore. The datastore supports querying data using a SQL query. SQL querying is beyond the scope of this document, but tutorials and example queries are widely available online. For more information about the CKAN Datastore, see: https://docs.ckan.org/en/2.8/maintaining/datastore.html.

1 Example Query

On the resource page for tabular data, you will see a green Data API button at the top-right. Click this to see examples of querying in action. Querying can be done from your web browser, or from R, Python, or any other programming language that supports HTTP requests and JSON. Results from queries against tables in the datastore are provided in the JSON format (see: https://en.wikipedia.org/wiki/JSON).

Each file uploaded to a dataset in CKAN is called a *resource*, and each resource has a unique resource ID autogenerated by the system. The resource ID appears at the end of the URL for the resource page. For example: https://livinglakesckan.ddns.net/dataset/large-flow-table/resource/c730d8a6-997c-436c-8c9e-795f91588c0e

In this example, the resource ID is: c730d8a6-997c-436c-8c9e-795f91588c0e. When performing an SQL query, the resource ID is the SQL table name. In the database behind CKAN, the table "c730d8a6-997c-436c-8c9e-795f91588c0e" contains all of the data (rows and columns) that was pulled from the uploaded Excel file. Here is a simple example, to get the first 5 rows:

https://livinglakesckan.ddns.net/api/3/action/datastore_search_sql?sql=SELECT%20*%20from%20%22f332bdde-fdef-43aa-a108-8ede7d712812%22%20LIMIT%205

The query in this case (without the URL formatting) is SELECT * from "f332bdde-fdef-43aa-a108-8ede7d712812" LIMIT 5

To execute a query, append the full SQL query to the CKAN datastore URL, and open the URL with a web browser, or using a language such as R or Python.

The CKAN datastore URL is: https://livinglakesckan.ddns.net/api/3/action/datastore_search_sql?sql=

In general, you can run any SQL query on any table in the datastore. Note that table and column names must be double quoted. For a complete example using R, see this page: https://livinglakesckan.ddns.net/dataset/example-r-script