CERA GUI Usage

Revision History

Revision	Author	Scope
February-2017	DKRZ Data management	Public release

Contents

Introduction

The CERA graphical user interface (GUI) is a powerful visual front-end to find and retrieve relevant datasets from within a huge collection of scientific datasets related to climate modelling. CERA is a long-term archive operated by DKRZ. It is certified according to the rules of the Data Seal of Approval (DSA) and is the technical basis for the World Data Centre of Climate (WDCC) which is as well hosted by the DKRZ.

Datasets are described with rich metadata elements following the CERA metadata standard. These medata elements allow users to find intended data objects and to download them if desired.
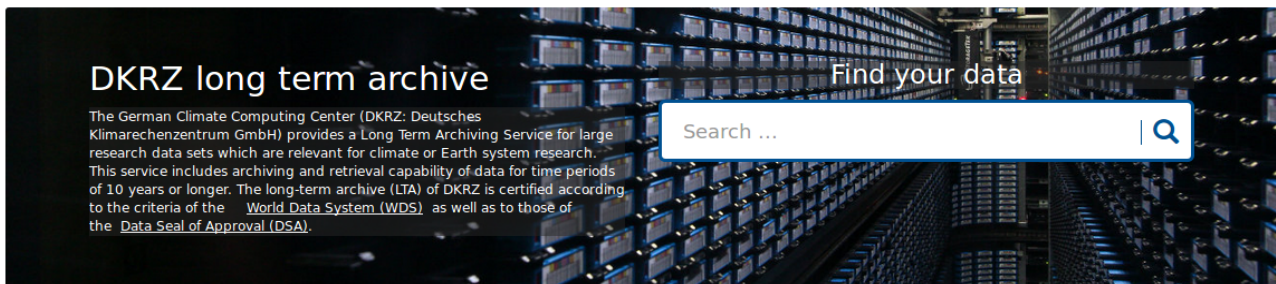
Intended Audience

Users of the CERA graphical user interface.

Interface

The CeraSearch portal is found at https://cera-www.dkrz.de/WDCC/ui/cerasearch/q. It is a graphical, web-based tool which is designed to be self-explanatory. The only required prerequisite to use the web interface is that JavaScript is enabled in the used browser.

In the CERASearch home page (see Figure 1), text can be entered into the free text field 'Find Your Data' in order to perform a full text search over the whole catalogue. You may just leave the text field empty and hit enter or click the magnifying glass in order to search for all available data. This leads you to a result page (Figure 2) where all hits found are listed on the right panel.

Figure 1: Landing page

There are several ways to find metadata records in CeraSerach, and powerful options to browse and filter search results are provided. These include the following:

1. Free text search over the full text body
2. Faceted search, i.e. selecting values of a metadata field or property
3. Filter by location or time, i.e. search for all datasets, that cover a chosen region or time period

These search requests can be combined and executed in one go.

A more detailed overview of all features of CeraSearch will be provided in the chapter *Search* at page 6.

Figure 2: CeraSearch interface

Browse

The best way to get an overview of the content of the archive is to follow the hierarchical structure implemented in CERA. Data in CERA is organized in multiple layers with a hierarchical structure. These layers are: Projects, Experiments, Dataset groups and Datasets. The Project, Experiment and Dataset groups contain the overall descriptions of the data. The datasets contain the description of the data and the climate data themselves.

Besides these layers, Additional infos are a fifth entry type which do not follow the strict hierarchy but can be connected to any entry. An Additional info is a compilation of documents or plots enhancing further understanding of the datasets.

As shown in Figure 3 you will see the entry type of a search result in the corresponding head bar.
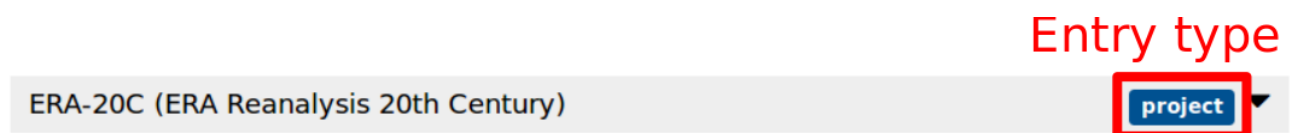
Figure 3: Position of the layer in the head bar of a search result

The entry point to the hierarchical structure is the search facet 'Project' (see Figure 4). By selecting

4

one project from the list you will get a list of all entries connected to this project.
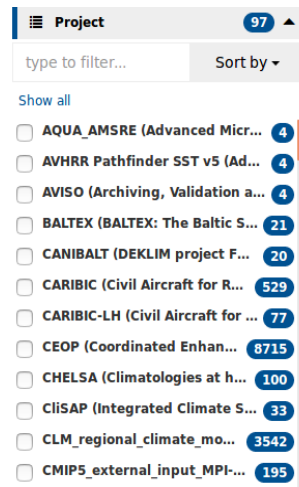
Figure 4: 'Project' facet

For example, you may select the project 'IPCC-AR5_CMIP5' which contains over 800,000 entries (including the Project itself). Next you may be interested in the data connected to an experiment (e.g. 'cmip5 output1 ICHEC EC-EARTH'). In order to find the subordinate entries of the experiment you have to open the corresponding search result by clicking its head bar (see Figure 5). At the bottom of the opened search result you may see its hierarchy. The right-most entry in this bar is the selected entry itself. Click the magnifying glass to perform a search for all subordinate entries of this experiment. If you are interested in a particular entry you may also click the "i" icon next to the magnifying glass in order to get detailed information.
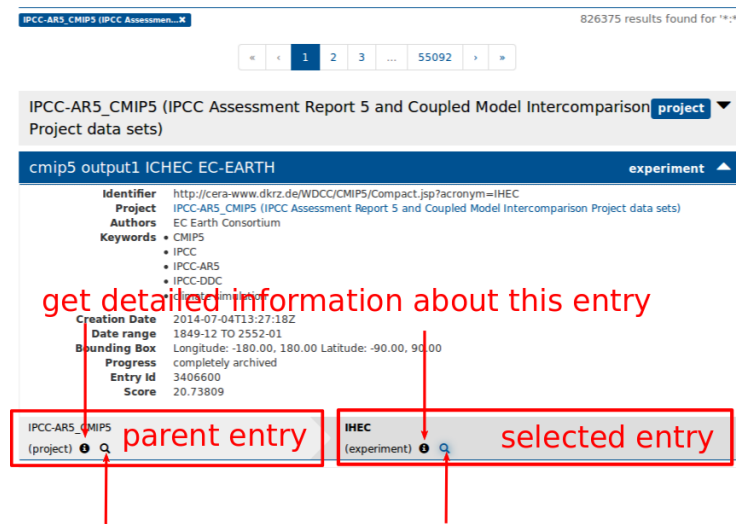
Figure 5: Information view of a search result

You can browse CeraSearch as described until you reach the level of datasets which do not have any

subordinate entries but are connected to actual data.

Please see the *CERA2 Metadata Submission Guide* to learn more about the hierarchical structure of CERA2 Metadata: http://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf

Search

Free Text Search

A full text search will be performed by simply typing the keywords you are interested in into the text field and hitting return. If you leave the text field empty, you will search for the wildcard "*:*" instead, which covers all available data in CERA.

Example: If you are interested in precipitation data from the CMIP5 project you may just type '*cmip5 precipitation*' to get the corresponding results. All results in the shown result set contain both terms (cmip5 AND precipitation) somewhere in their metadata. However, it is possible to refine your search by using the search operators described at page 18.

Please note that in the search result a subset of the indexed metadata is shown. Therefore, it may not be always immediately obvious why a result is displayed for a particular search.

Faceted Search and Filters

The faceted search provides you with options to filter your search by choosing 'facets'. This tool may help you to narrow down the search results for your specific search demands.

Furthermore, CeraSearch provides the opportunity to filter out datasets by space or time.

You may find all available search facets and filters at the left panel of the search interface.

Please note, that by enabling facets and filters you will receive search results which fulfill all of the requirements defined by the filters (AND-Operator).

Spatial filter

By clicking the green button in the world map widget, you may enable/disable the geospatial filter. In order to drag or resize the selected area click and move one of the corners, or the cross in the middle of the red bordered area.
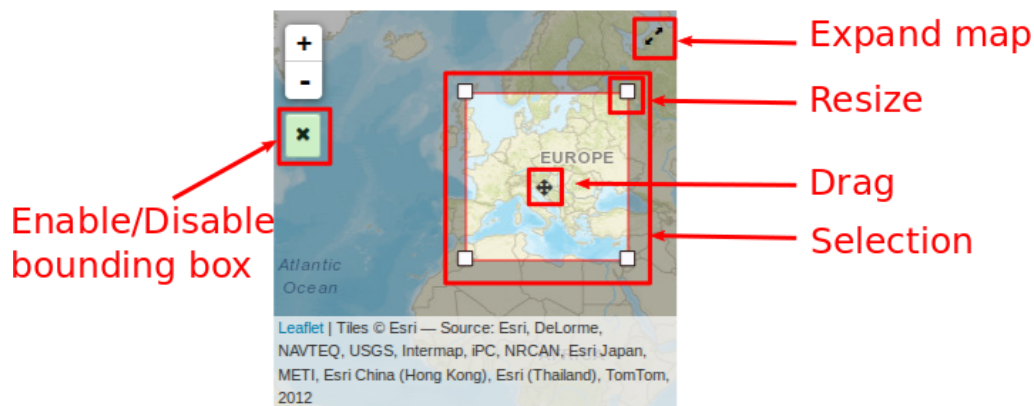
Figure 6: Map widget

If you prefer to enter exact coordinates, you can do so after unfolding the widget labeled 'Bounding box'. You may also refine your search by selecting one of the search operators described at page 18.

Do not forget to apply your selection by clicking the magnifying glass at the search bar or the "Apply"-button in the unfolded "Bounding Box" widget.

Temporal filter

With the search widget 'Temporal Coverage' you can select a time period the data are related to. This filter works the same way as the spatial filter described above and provides the same search operators.

In order to search for a date before 1 AD just enter a negative number.

Keep in mind that not all data provide a real date information but instead refer to 'model runtime'. For example, if one dataset has a temporal coverage from year 1 to 140, this may mean that the dataset has a runtime over 140 unspecific years and not the particular years 1 to 140 AD.

Do not forget to apply your selection by clicking the magnifying glass at the search bar or the "Apply"-button in the unfolded "Temporal Coverage" widget.

Facets

All remaining widgets at the left panel are so-called facets. Facets are categories by which you can classify search results. When you open one facet widget and select one of the subcategories, the search is narrowed, and a new search limited to this subcategory is performed. The number in the header of each widget shows you the amount of subcategories for that particular facet. The number behind each subcategory represents the count of search results which will be shown, if you enable this subcategory.
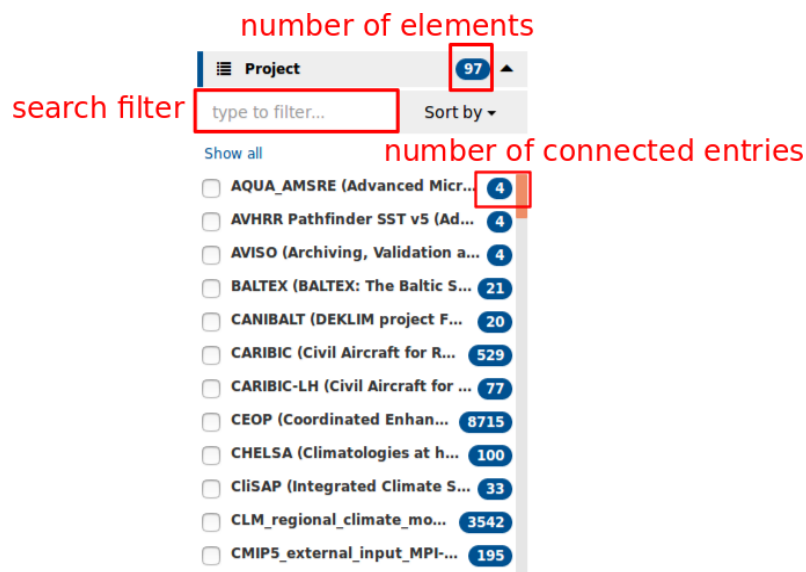
By typing something into the text field labeled 'type to filter…' you may search for a particular value in the selected facet. The widget will be updated automatically.

If you need an overview of all possible subcategories, you may click the 'show all'-button to display a popup menu containing all subcategories for the selected facet.

Furthermore, you can sort all facets alphabetically or by their amount of hits.

Search Results

The datasets found to fulfill the search criteria are shown on the result page on the right side of the portal. You can unfold them to get a description and a short overview by clicking the gray head line (Figure 7).
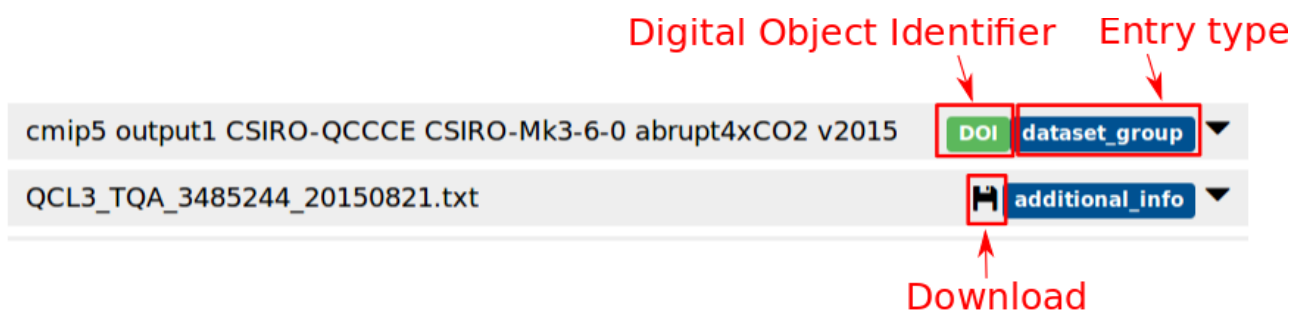

Figure 7: Head lines of two entries

 The labels at the right hand side of the head line are clickable as well. There are three kinds of labels:

    - **Entry type** – The labels with blue background are type information following the CERA data model. Please see the chapter about the CERA data model to get more information about them.

By clicking this label you will be directed to the metadata page for this particular search result.

   - **DOI** – All data toi which a digital object identifier (DOI) is assigned are marked accordingly with a green 'DOI' label.

Please note, that only 'experiments' (according to the CERA data model) and dataset groups can have a DOI.

   - **Download** – A floppy disc icon at the head line indicates downloadable data. These include only additional infos and datasets. If an additional info or a dataset is missing the floppy icon, then this data is metadata only. By clicking the floppy disk, you will be forwarded to the download form.

Please note, that the CeraSearch Web Interface has no knowledge about your download permissions. Even if there is a floppy disc icon, the download form may disallow you to download the data, e.g. if the dataset in question requires special access clearance.

Find

This chapter describes a set of typical use cases and helps to solve common problems.

Find a DOI

Use case: Search for data for the DOI *doi:10.1594/WDCC/CMIP5.CQMKhmv2015*

Just enter the DOI into the search field and press at the magnifying glasses. The result will look like in Figure 8.
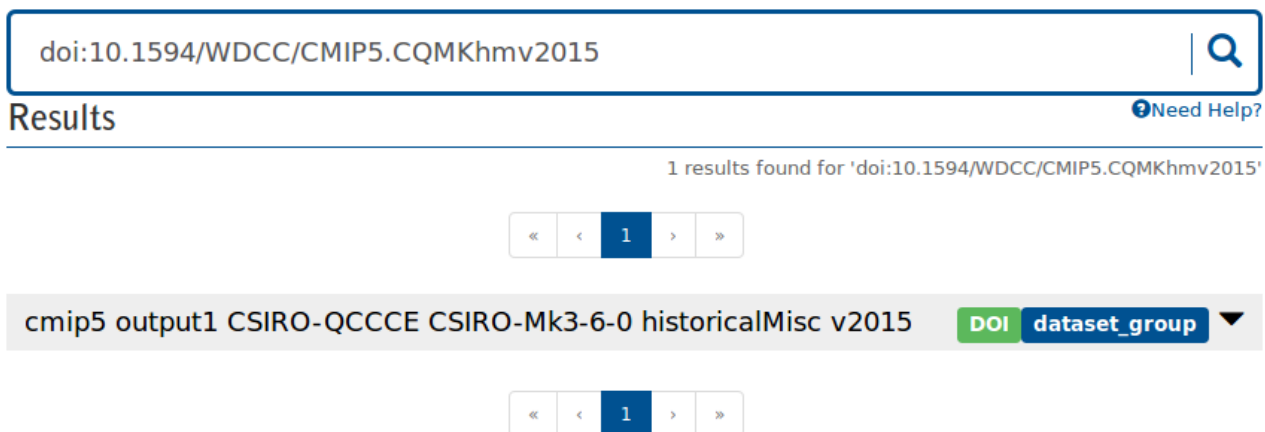

Figure 8: Search for a DOI

Find data for a project

Use case: We are looking for all available data of the project CHELSA.

1   Select project

    1.1   Untoggle the 'Project' facet

    1.2   Enter 'Chelsa' in the filter form. Only the CHELSA Project should be visible now.

    1.3   Pick 'CHELSA'

2   Select datasets only

    2.1   Untoggle the 'Type' facet

    2.2   Pick 'dataset'. We are done.

Figure 9: All datasets of the CHELSA Project

Find data for an experiment

Use case: We are looking for data of the project IPCC-AR5_CMIP5 and the experiment MPI-ESM-MR which contains monthly data with the topic sea_water_salinity.

1. Find the experiment

    1.1.   Enter 'MPI-ESM-MR' in the search field.

1.2. So far we are only interested into experiments. We select 'experiment' in the 'Type' facet (see Figure 10).

1.3. Since there a only 10 results left, it is easy to find the experiment now. If you are unsure, you may select select the project 'IPCC-AR5_CMIP5' in the 'Projects' facets. This leads to only one result: The experiment we are looking for.

2. Find data

2.1. Open the information overview of the experiment and start a hierarchical search for its children by clicking the magnifying glass (see Figure 11). Please note, that all previously selected facets and search terms will be replaced by the hierarchical search.

Figure 10: Search interface after step 1.2

2.2. Narrow down the result set by selecting the topic name 'sea_water_salinity' in the 'Topic Name' facet. You may notice that for this topic name there is no other aggregation than monthly mean. This result set contains 134 datasets. We are done.

Figure 11: Experiment overview

Download

Depending on the type of data you want to download there are some differences in the displayed download forms.

Datasets containing time series data allow selection of time spans either by specifying start- and end-date, if the data is available in a calendrical format, or by specifying start- and end-record otherwise.

Figure 12: Download form for time series data

Archived files

The download form for datasets that contain archived files displays a list of available files and the corresponding sizes. You can either view file properties, download a single file, download all files or create a comma-separated-list of files to use externally. If you choose to download all files they will be packaged into a single zip-archive during download.

Figure 13: Download form for archived files

DKRZ project files

On the download form for datasets containing DKRZ project files you will find a file list including file sizes and checksums (md5). You can access additional file properties, download individual files or export the list in .csv format.

Figure 14: Download form for DKRZ project files

Additional info

Additional info data objects contain one or more documentation file(s), i.e. PDFs etc. You can download the files separately or all packaged into one zip-archive. If there is only one file available the download will start immediately without displaying the form first.

Figure 15: Additional info download form

REST API

In addition to the web frontend we also provide a REST API in order to enable our users to define custom search requests. We are using Apache Solr 6 (https://lucene.apache.org/solr/) as our search engine. You may use most of the query functionalities listed in the Solr Reference Guide (http://archive.apache.org/dist/lucene/solr/ref-guide/apache-solr-ref-guide-6.4.pdf). A small subset of the most useful Solr parameters is listed in the appendix at page 19.

Example: You may start with a very basic GET-Call by calling the following URL:

https://cera-www.dkrz.de/WDCC/ui/cerasearch/solr/select?
indent=on&q=CMIP5&start=0&rows=10&wt=json

The result of this call will be a JSON-File (wt=json - JSON is default; XML is also possible) containing 10 hits (rows=10) out of the found results for your search query which matches the term 'CMIP5' (q=CMIP5).

Fields

A field is a chunk of information about one result. For example, the field which the Web interface uses as the result title in the head bars is entry_name_s. You may use fields e.g. in conjunction with the "q=" parameter to search for results which contain a specific value in a particular field.

Example: If you want to search for a specific result which contains the string "H2O_Climate_v1.0" in its title, you may use the following URL:

https://cera-www.dkrz.de/WDCC/ui/cerasearch/solr/select?
indent=on&q=entry_name_s:H2O_Climate_v1.0&start=0&rows=10&wt=json

Please note that all fields containing the suffix "_s" or "_ss" like entry_name_s are so-called string fields. Those are case-sensitive, and you may only search for substrings by using the wildcard character "*".

The most useful fields will be described in the appendix at Page 20.

Use Case

A possible use case for the REST API might be the instrumentation of Jblob. By using the REST API with curl, sed and grep you may create simple download scripts.

Example: Let's imagine you want to download all files of the experiment with entry acronym CQMKc1v2015 of the CMIP5 project. For this use case you may use the following query the get all entry_acronyms:

https://cera-www.dkrz.de/WDCC/ui/cerasearch/solr/select?
&wt=json&indent=on&q=hierarchy_steps_ss:CQMKc1v2015&start=0&fl=entry_acronym_s&f

[q=entry_type_s:dataset&rows=1000](#)

Since we don't know the amount of result we will get, we are using the maximum size of 1000 rows (the actual amount of result is much smaller). After receiving the result set via curl, you may parse it via grep and sed in order to get a list of valid jblob calls. The complete script generation may look like this:

```
curl -sk 'https://cera-www.dkrz.de/WDCC/ui/cerasearch/solr/select?
&wt=json&indent=on&q=hierarchy_steps_ss:CQMKc1v2015&start=0&fl=entry_acronym_s&f
q=entry_type_s:dataset&rows=1000' | grep entry_acronym_s\": | sed -e
's/\"entry_acronym_s\":\"\(.*\)\"\}.*/jblob --dataset \1/'
```

Support

If you need any further assistance please contact data@dkrz.de.

Appendix

Full text search operators

Operator	Example	Description
OR	cmip5 OR cordex	Will show results which contain at least one of the search terms
AND	cmip5 AND cordex	Will show results which contain both search terms; AND is the default operator (i.e. '*cmip5 AND cordex*' is equivalent to '*cmip5 cordex*')
NOT	NOT cmip5	Will show all results which do not contain the search term

It is also possible to combine these operators in order to create more complex search queries. For example, if you are interested in precipitation data which is not connected to cordex or cmip5 you may search for '*precipitation AND NOT (cmip5 OR cordex)*'.

Temporal and spatial search operators

Contains	Shows only results which contain your selection (default).
Intersects	Shows only results which intersect with your selection. You will get the largest amount of results by choosing this operator.
Within	Shows only results which are within your selection. You will get the smallest amount of results by choosing this operator.

List of available facets

Project	A CERA-project is a scientific research activity consisting of experiments. It enables data depositors to organize their

	collections of data from a number of potentially different data sources. Examples: IPCC-AR5_CMIP5, HZG, CliSAP.
Keyword	Words or phrases used to further describe an entry.
Topic Name	Kinds of measurements/physical quantities represented by the data. A keyword list exists (e.g. sea level).
Aggregation	Description of the time method used to derive data that represents the topic values e.g. 6hr, 3hr, monthly mean.
Type	Type of the search result according to the CERA data model
Code Type	Naming of model-specific variable code lists. A model variable code list includes code numbers, internal names and the description of the physical quantities with units.
Institute	Institutes which are connected to the entry in some way (data provider, creator etc.)
Format	File format of the data connected to datasets and additional infos
Progress	The state of the data set in the CERA database (e.g. complete, in work, planned).

Overview of available SOLR parameters

Parameter	Description
&q	'Query' is a mandatory parameter in order to get results from your REST call. The most simple way to use 'q' is to write your search term right after the parameter, e.g. &q=CMIP5. You may also look for your search term in a particular field, e.g.: &q=entry_name_s:cmip5 This query will look for results with their entry names exactly matching 'cmip5'. The most useful fields are described in the chapter below. In contrast to a simple full text search, using fields in the query requires you to know the exact value for this field (case-sensitive!). The example shown above will lead to no results, since there is no result which has the entry name 'cmip5'. However, you may use wildcards (*). Therefore, a functioning example of the query above which look for all results starting with cmip5 would be the

	following: &q=entry_name_s:cmip5* Please note that you have to You may also use all set operators mentioned at page 18, e.g: &q=entry_name_s:CMIP5* OR entry_name_s:ETH*
&fl	You may use fl in order to reduce the number of shown fields, you get by the REST call. For example, by adding '&fl=entry_name_s' the only information (field) you will receive is the name of each result. To define multiple fields just add them after a comma, e.g. fl:entry_name_s,entry_type_s. Note that using fl has no influence on the search query itself, i.e. Solr will also look for your search terms in all fields you set to invisible due to fl.
&start	This parameter is used to paginate results from a query. When specified, it indicates the offset in the complete result set for the queries where the set of returned documents should begin
&rows	This parameter is used to paginate results from a query. It specify the maximum number of documents from the complete result set to return to the client for every request. You can consider it as the maximum number of result appear in the page. The default value is "10", which is used if the parameter is not specified. The maximum value is limited to 1000.

List of available SOLR fields

entry_name_s	Title of the result
entry_acronym_s	Acronym of the result, which we use internally in CERA.
entry_type_s	Data type based on the CERA specification. Possible values are: - project - experiment - dataset_group - datasets - additional_info
hierarchy_steps_ss	List of all entry acronyms of superordinate results of this result
data_size_l	Complete size of this result. Only available for additional_infos and datasets.
text	This special field contains all available information about one particular

	result. If you don't specify any field in the q-Parameter, your search will be performed on _text_ by default. It is the only field which does not require you to use any exact search term. Consequently, it is also not case-sensitive.

Sort order

The sort order is based on the following factors:

- Type – Projects > Experiments > Dataset Groups > Datasets/Additional Infos

- DOI – Experiments with DOI have a better score

- Recency – Newer results have a better score

- Search term – Results with the search term appearing in their title have a higher priority. If you use the OR-Operator, results fulfilling all search conditions will have a higher priority than those which fulfill only a subset.

- Frequency of search terms – Results which contain the search term multiple times have a higher priority.

- 'Uncommon' search terms – If you use the OR operator to combine a 'common' term (e.g. 'temperature') with an 'uncommon' term (e.g. BALTEX), results containing BALTEX will have a higher priority since the uncommon word seems to be more specific.