Contents

1 CITREC Database – Documentation

This document describes the structure of the CITREC database and gives a brief tutorial on how to use the demo database via the phpMyAdmin web interface.

1.1 Database Overview

Figure 1 presents the relational schema of the CITREC database.

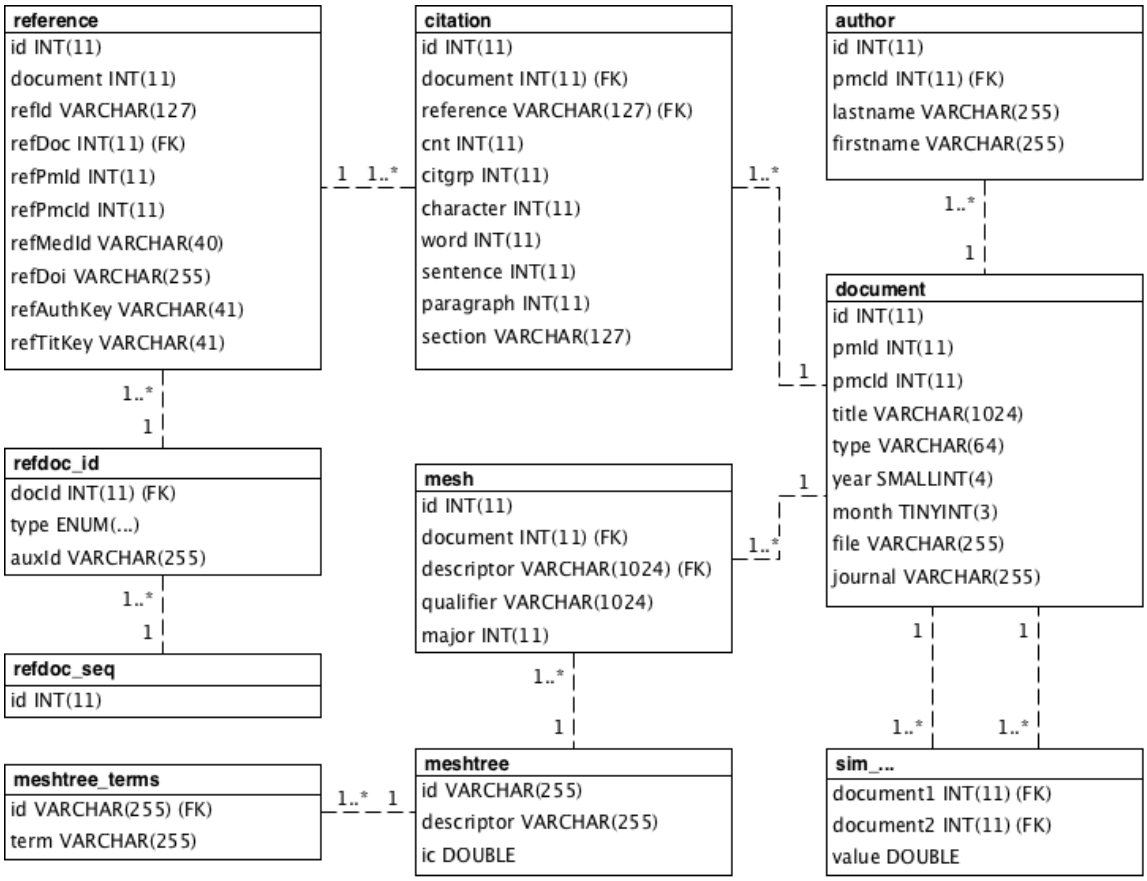

Figure 1: CITREC Database Schema

The *document* table stores the main metadata for articles contained in the test collection such as title, PubMed/PubMed Central identifier (PmId, PmcId) and year of publication. Due to the $1{:}n$ nature of the relation between documents and authors, author data are stored in a separate *author* table. Data about documents that articles reference in their bibliography are stored in the *reference* table. Multiple identifiers such as PmId, Digital Object Identifiers (Doi) and automatically generated keys from author- and title substrings are assigned dependent on their availability for matching and integrating references parsed from several articles. The *refdoc_id* and *refdoc_seq* tables are used internally by the data preprocessing tools provided in the CITREC framework for performing reference disambiguation. The *mesh* table stores MeSH descriptors assigned to individual articles in the test collection, while the *meshtree* and *meshtree_terms* tables (only available for the TREC

Genomics part of the test collection) store the complete MeSH taxonomy. Tables having the prefix *sim_* store pre-computed similarity scores obtained by applying particular similarity measures.

1.2 Table Details

This section describes the attributes of tables in the CITREC database.

1.2.1 document

Attribute	*Description*
id	Internal document identifier (currently unused)
pmId	PubMed identifier (if available)
pmcId	PubMed Central identifier (if available)
title	Document title
type	Article type, e.g. research-article, review-article, errata or comment
year	Year of publication
month	Month of publication
file	Relative path to the source file of the document
journal	Abbreviated title of the journal the article was published in

1.2.2 author

Attribute	*Description*
id	Internal identifier for an author occurrence (currently unused)
pmId	PubMed identifier of a document written by the author
lastname	The author's last name
firstname	The author's first name

1.2.3 citation

Attribute	*Description*
id	Internal identifier of the citation (currently unused)
document	PubMed identifier of the document containing the citation
reference	Identifier of the reference that the given citation links to
cnt	Consecutive number of the citation within the document's full text
citgrp	Consecutive number of the citation group that the citation belongs to within the document's full text
character	Character count at which the citation occurs within the document's full text
word	Word count at which the citation occurs within the document's full text
sentence	Sentence number in which citation occurs within the document's full text
paragraph	Paragraph number in which citation occurs within the document's full text
section	Section number in which citation occurs within the document's full text

1.2.4 reference

Attribute	*Description*
id	Internal identifier of the reference (currently unused)
document	PubMed identifier of the document containing the reference
refId	Identifier of the reference (unique within the containing document only)
refDoc	Identifier of the referenced document
refPmId	PubMed identifier of the referenced document
refPmcId	PubMed Central identifier of the referenced document
refMedId	MEDLINE identifier of the referenced document
refDoi	DOI identifier of the referenced document
refAuthKey	Identifier of the referenced document generated from author name substrings
refTitKey	Identifier of the referenced document generated from title substrings

1.2.5 refdoc_id

Attribute	*Description*
docId	Identifier (same as in refdoc_seq) of a referenced document
type	Type of the auxiliary identifier (pm, pmc, medline, doi, title, authors)
auxId	Value of the auxiliary identifier

1.2.6 refdoc_seq

Attribute	*Description*
id	Auto-incremented identifier assigned to referenced documents

1.2.7 mesh

Attribute	*Description*
id	ID of the mesh descriptor assigned to a document (currently unused)
document	PubMed identifier of the document
descriptor	The MeSH descriptor assigned to a document
qualifier	The MeSH qualifier for the descriptor assigned to a document
major	Flag indicating if the descriptor (= 1) or the qualifier (= 2) are marked as major

1.2.8 meshtree

Attribute	*Description*
id	Identifier of a MeSH descriptor (includes information about the descriptor's position in the MeSH tree)
descriptor	The MeSH descriptor
ic	Information Content of the MeSH descriptor

1.2.9 meshtree_terms

Attribute	*Description*
id	Identifier of a MeSH descriptor (includes information about the descriptor's position in the MeSH tree)
term	Term of the MeSH descriptor

1.2.10 sim_…

Attribute	*Description*
document1	The **PubMed** identifier of the first document
document2	The **PubMed** identifier of the second document
value	Similarity score of the two documents. Some similarity measures are **not** symmetrical, i.e. the similarity between document1-document2 and document2-document1 is not identical. In those asymmetrical cases, document1 was the input for which similarities have been calculated.

2 CITREC Database – Tutorial

This section gives a brief tutorial on how to access and query the database using the phpMyAdmin web interface. The interface is available at:

http://46.4.84.169/phpmyadmin/?db=citrec_demo.

2.1 Login

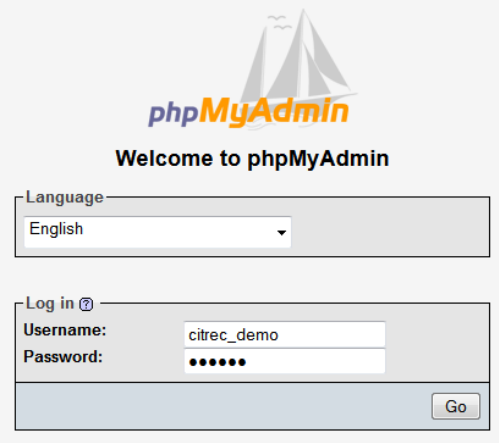

Figure 2: Login Screen

The login credentials for the demo system are:

Username: citrec_demo
Password: citrec

2.2 Table Overview

The main view after login presents an overview of the available tables (relations).

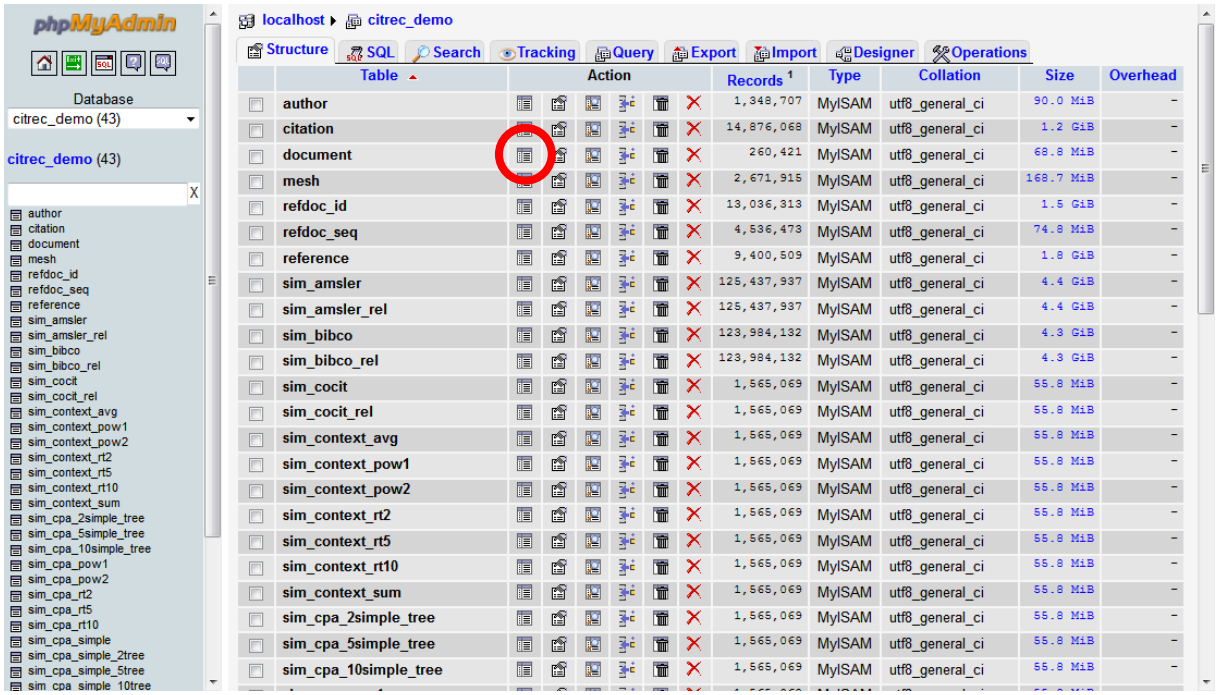

Figure 3: Table Overview Screen

Click on the “browse” button to explore the entries of particular tables. For our example, we browse the table *document* to have a look at the documents stored in the database.

2.3 Detailed Table View

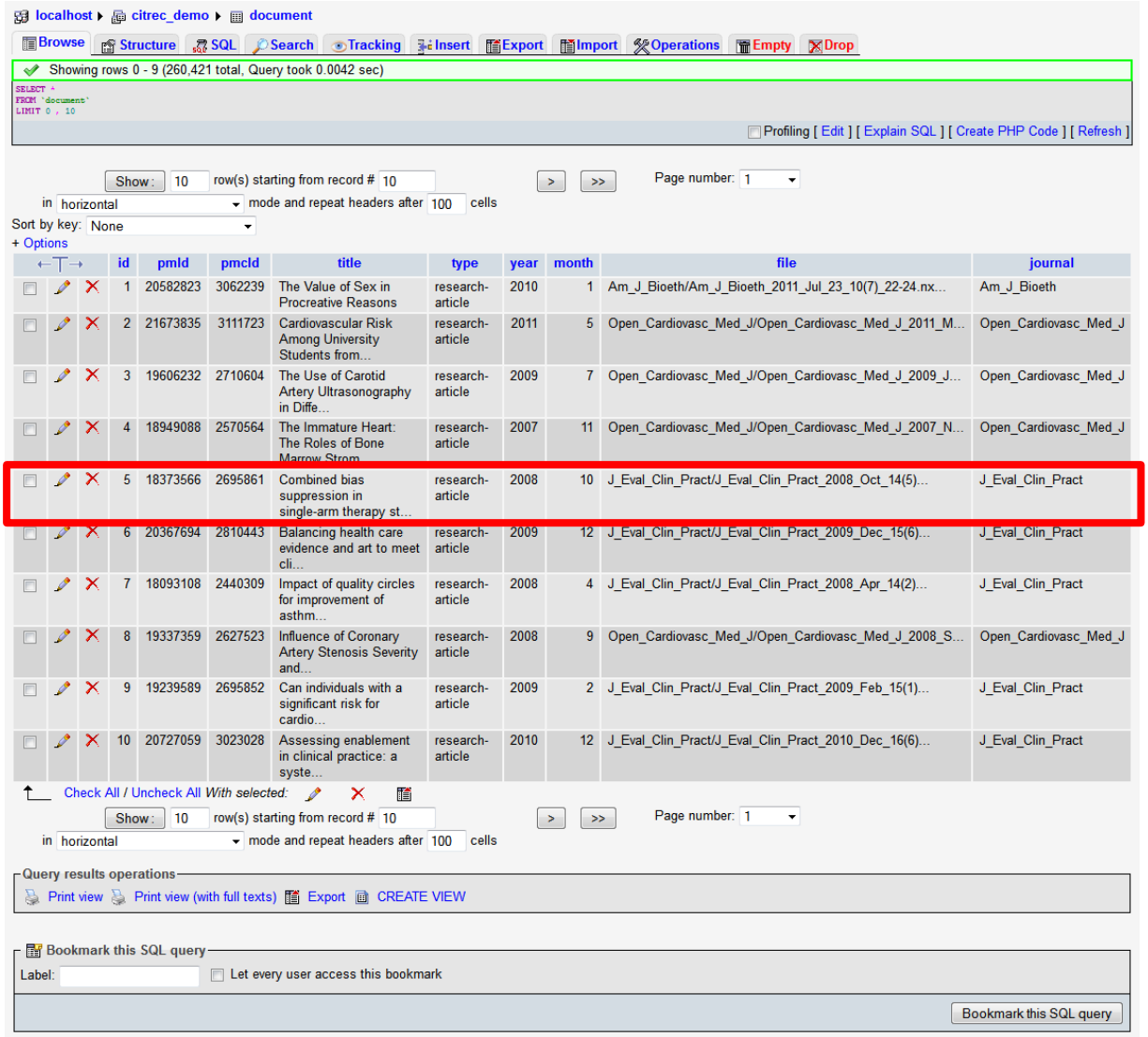

Figure 4: Detailed Table View

Let us assume we are interested in articles that co-cite the fifth document "Combined bias suppression in single-arm therapy studies". To do this, we need to copy the PubMed identifier (PmId) of the respective document as shown in Figure 5.

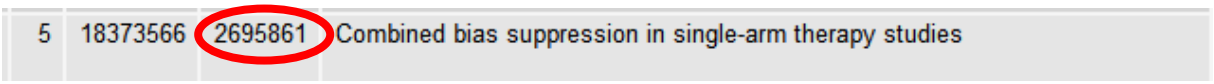

Figure 5: Copying the PmId of an Article

2.4 Switching Tables

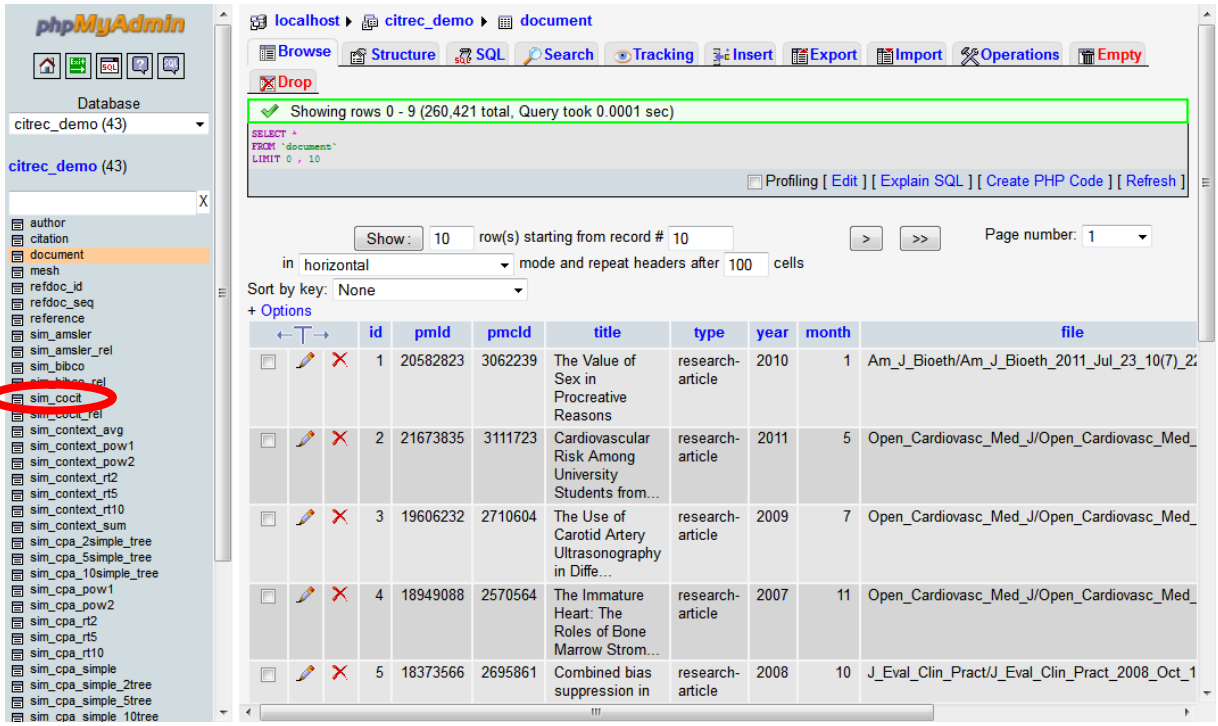

Figure 6: Switching Tables

Because we are interested in Co-Citation information, we have to switch to the table *sim_cocit* by clicking on the respective entry within the table overview panel on the left side of the screen (see Figure 6).

2.5 Searching for Table Entries

To search for specific entries contained in a table, click on the respective tab in the upper area of the screen (Figure 7).

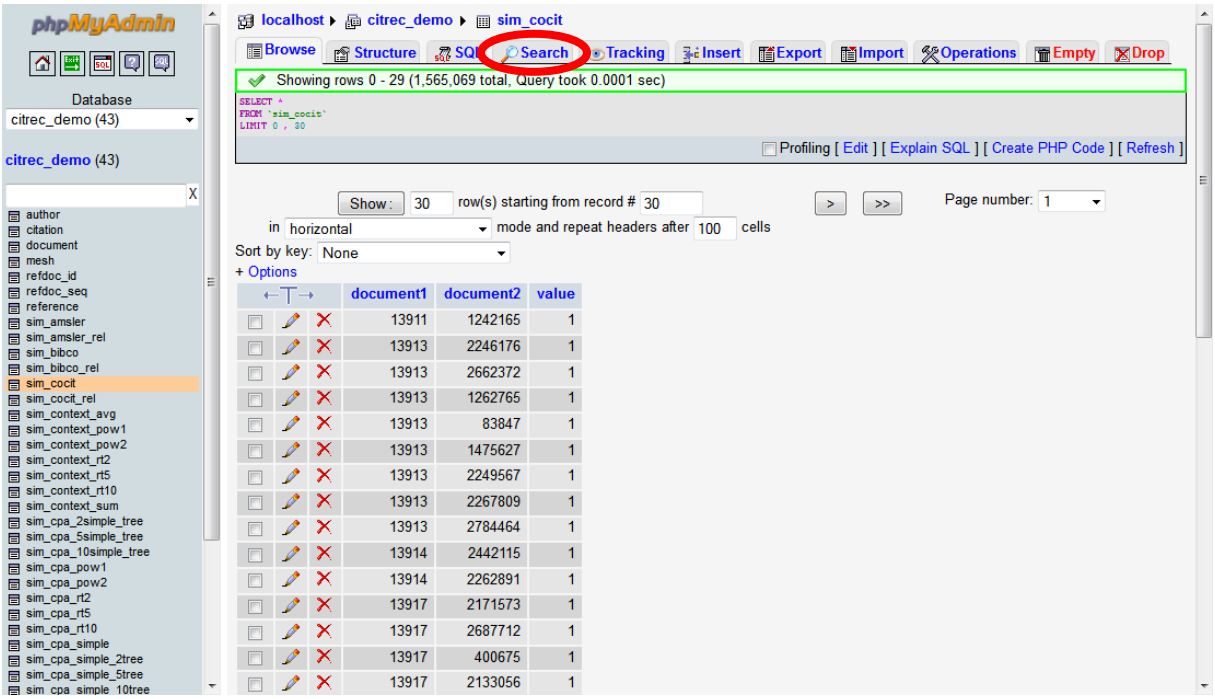

Figure 7: Navigating to the Search Tab

Paste the copied PmId into the field “document1” within the search form presented under the search tab and hit go (see Figure 8).

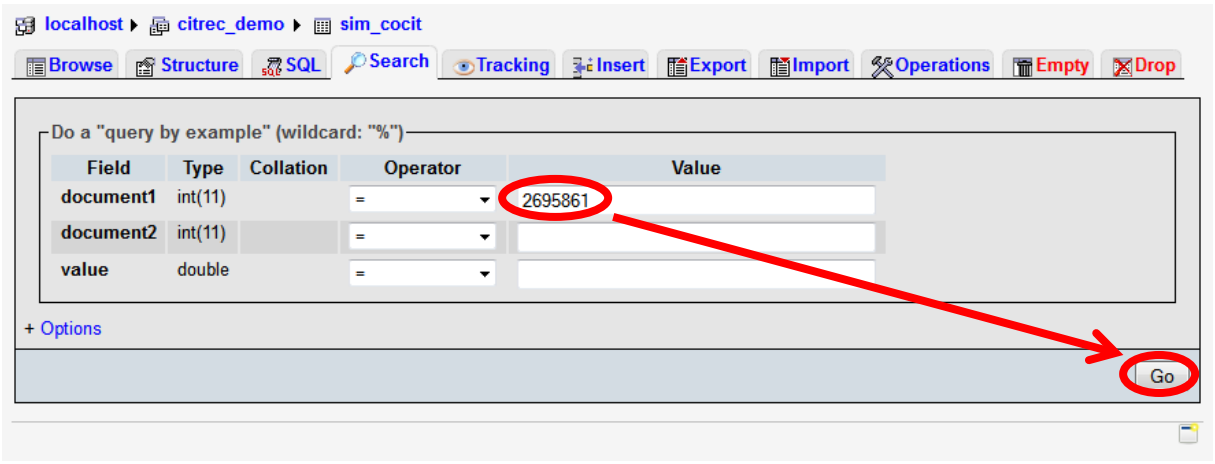

Figure 8: Searching

The results of the search show you all other articles that the article in question was co-cited with. “Value” indicates the number of times the respective articles were co-cited (Figure 9).

	document1	document2	value
	2695861	395833	1
	2695861	1397812	1
	2695861	1513220	5
	2695861	117123	5
	2695861	1800863	3
	2695861	1298309	1
	2695861	1764730	4
	2695861	1868723	3
	2695861	1876246	1
	2695861	1947979	1
	2695861	1863422	1
	2695861	2440731	1
	2695861	2713221	3
	2695861	2386850	1
	2695861	3004618	1
	2695861	2816246	1
	2695861	1924504	1
	2695861	2323398	2
	2695861	2934607	1
	2695861	2761176	1
	2695861	2570361	1

Figure 9: Search Results

For sorting the result set according to Co-Citation strength click on “value” (clicking once sorts entries in ascending, clicking twice in descending order).This way, the most similar documents can be displayed easily (see Figure 10).

	document1	document2	value
	2695861	1513220	5
	2695861	117123	5
	2695861	1764730	4
	2695861	1868723	3
	2695861	2713221	3
	2695861	1800863	3
	2695861	2323398	2
	2695861	2934607	1
	2695861	2761176	1
	2695861	1924504	1
	2695861	2816246	1
	2695861	3004618	1
	2695861	395833	1
	2695861	2386850	1
	2695861	2440731	1
	2695861	1863422	1
	2695861	1947979	1
	2695861	1876246	1
	2695861	1298309	1
	2695861	1397812	1
	2695861	2570361	1

Figure 10: Sorted Search Results

2.6 Advanced Queries

If you are familiar with the SQL query language, you can also use more complicated ad-hoc queries. Click on the SQL tab (Figure 11) to open the SQL Query Editor for creating your SQL statement.

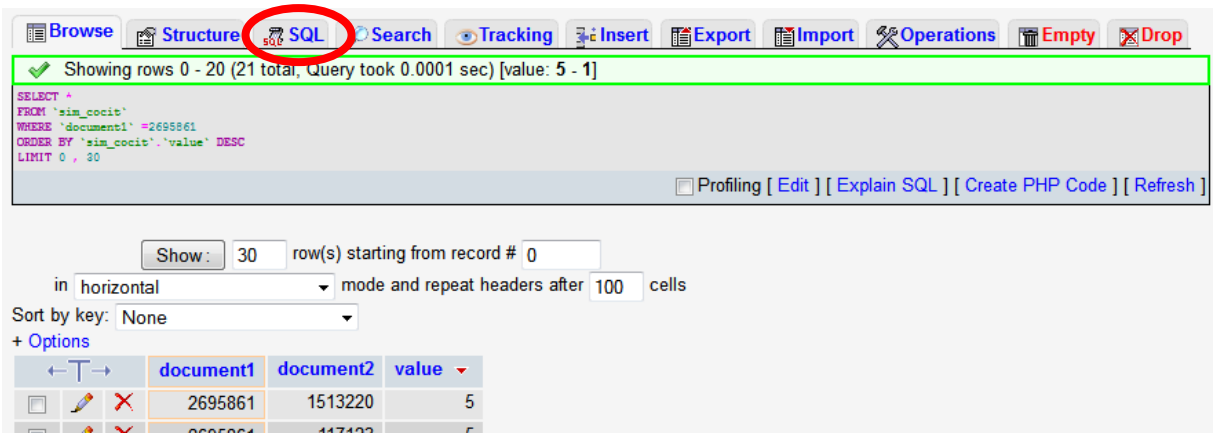

Figure 11: Using the SQL Query Editor

As an example for the plurality of interesting information that can be readily obtained from the database, we print the distribution of Co-Citation strengths by using the statement in Figure 12.

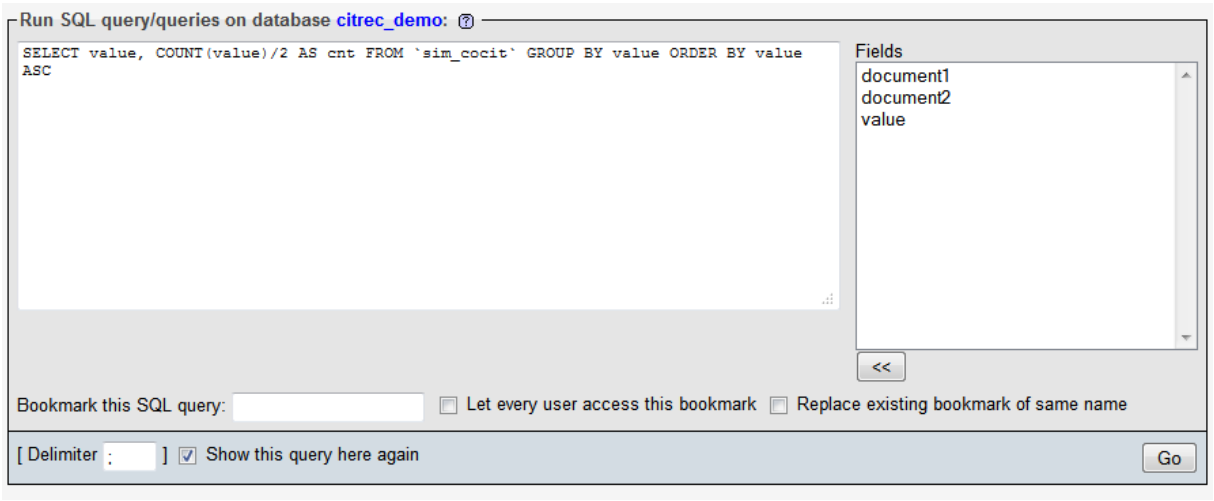

Figure 12: Ad-hoc Query Example

value	cnt
1	682527.0000
2	64185.5000
3	18091.0000
4	7616.0000
5	3756.0000
6	2126.0000
7	1281.0000
8	819.0000
9	510.0000
10	386.0000
11	267.0000
12	205.0000
13	145.0000
14	120.0000
15	89.0000
16	71.0000
17	50.0000
18	54.0000
19	34.0000
20	25.0000
21	19.0000
22	17.0000
23	20.0000
24	17.0000

Figure 13: Ad-hoc Query Results

Feel free to browse and analyze the data to see what interesting details you can reveal from it. If you encounter any questions or problems, please contact us at: team@sciplore.org