## Features/Counts Testing

In general the numbers from Solr will be slightly lower due to the changes in the document set since the snapshot for HotHouse was taken. Other minor variations may exist due to the transformation of content not being exactly right, stemming variations, or configuration/mapping issues within Solr. For example, we did notice a CIP mapping issue with author first name. The scope feature is not implemented out of the box in Solr and would require custom development or a third-party plug-in such as QPL.

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Scopus Count** | **SOLR Count** | **Query/Notes** |
| CORE1 | 1 | 1 | doi(10.1109/lawp.2005.850798) |
| CORE2 | 2 | 2 | eid(2-s2.0-84861409323 OR 2-s2.0-0032977045 OR 2-s2.0-84871047622) |
| CORE3 | 1334 | 1315 | refeid(2-s2.0-0034824859) |
| CORE4 | 14 | 15 | au-id(55177469200) |
| CORE5 | 122599 | 121545 | af-id(60026851) |
| CORE6 | 2 | 0 | scopus-id(33749522827 OR 33750354968) |
| CORE7 | 1927 | 1927 | issn(2090-4649 OR 0301-0538 OR 004-2675-X) |
| CORE8 |  | 0 | fulltext-eid(1-s2.0-S0169409X13002275) |
| CORE9 | 15 | 15 | srcid(24527) AND PUBYEAR = 2012 |
| CORE10 | 4920871 | 4779881 | all(blood) |
| CORE11 | 3524700 | 3499039 | title-abs-key-auth(blood) |
| CORE12 | 3523591 | 3498027 | title-abs-key(blood) |
| CORE13 | 1383826 | 1383796 | abs(blood) |
| CORE14 | 209717 | 207911 | affil(oxford). |
| CORE15 | 863 | 846 | chemname(rhamnose) AND indexterms (polysaccharide). |
| CORE16 | 1021 | 1002 | key(rhamnose AND polysaccharide). |
| CORE17 | 4145698 | 2388514 | ref(engineering). ML had similar results. Since we used the ‘storage’ format and not the ‘syndication’ formation for the XML, it’s possible that additional information was included in the ‘syndication’ format. |
| CORE18 | 60734 | 60580 | abs(blood) AND abs(cancer) |
| CORE19 | 664325 | 656006 | (abs(brain) AND NOT abs(cancer)) |
| CORE20 | 1611025 | 1597394 | (abs(brain) OR abs(cancer)) |
| CORE21 | 3308 | 2960 | abs(brain PRE/5 cancer). SOLR Surround Query Parser will not stem tokens. QPL could address this limitation. |
| CORE22 | 5045 | 4499 | abs(brain W/5 cancer). SOLR Surround Query Parser will not stem tokens. QPL could address this limitation. |
| CORE23 | 486 | 141 | title (practice W/5 "letter"). SOLR Surround Query Parser will not stem tokens. QPL could address this limitation. |
| CORE24 | 600 |  | title(neuro\* W/3 letter). SOLR errored on too many terms. . Would require custom development or QPL. |
| CORE25 | 24767 |  | title(neurological) AND NOT title(neurological W/5 letter). Could not construct query in SOLR (parse error). Would require custom development or QPL. |
| CORE26 | 754 | 141 | title((neurological OR letter) w/5 practice). SOLR Surround Query Parser will not stem tokens. . Would require custom development or QPL. |
| CORE27 | 3 | 3 | auth(Rumiá Arboix J). |
| CORE28 | 103987 | 103320 | title(blood) AND pubyear AFT 2003 |
| CORE29 | 23933106 | 23650409 | pubyear AFT 2000 AND pubyear BEF 2014 |
| CORE30 | 3727 | 3696 | abs("heart attack") |
| CORE31 | 21 | 13 | abs({heart-attack}) |
| CORE32 | 43462156 | 43082537 | all("the") |
| CORE33 | 16330 | 16268 | title ("the end" ) |
| CORE34 | 96730 | 95800 | abs(thio\*e) |
| CORE35 | 5325 | 3052 | abs(thio?e). Unclear why Solr has is 2K less than Scopus. |
| CORE36 | 2924 | 2787 | title("\*snake venom") |
| CORE37 | 2924 |  | title (\*snake pre/0 venom). Solr has a praser error. Would require custom development or QPL. |
| CORE38 | 65626 |  | author-name(smith,j). Not implemented. Would require custom development or QPL. |
| CORE39 | 276 |  | author-name(smith,john). Not implemented. Would require custom development or QPL. |
| CORE40 | 81374 |  | affil(affilorg (university california) AND affilcity(san diego)). Not implemented. Would require custom development or QPL. |
| CORE41 | 22488 |  | ref(sanger nicklen DNA). Not implemented. Would require custom development or QPL. |
| CORE42 | 6178 |  | ref(refauth(darwin) AND refsrctitle(species) AND refpubyear IS 1859). Not implemented. Would require custom development or QPL. |

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Scopus Count** | **SOLR Count** | **Query/Notes** |
| AUTH1 | 1 | 1 | au-id(10038760900) |
| AUTH2 | 11686 | 11479 | auth-last-name(smith) AND auth-first(j\*). |
| AUTH3 | 7021 | 4599 | auth-last-name (EQUALS(smith)) AND auth-first (STARTS-WITH(j)). Mapping issue for Solr in the CIP. |

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Scopus Count** | **SOLR Count** | **Query/Notes** |
| AFFIL1 | 1 | 1 | af-id(60030162) |
| AFFIL2 | 348 | 17835 | affilcity(london). Likely attributed to differences in publication counts as well as identification of root nodes. MarkLogic and SOLR had same count (25,847) when all filtering was removed. |
| AFFIL3 | 1 | 4 | affil("university of york"). Likely attributed to differences in publication counts as well as identification of root nodes. MarkLogic and SOLR had similar counts (9/6) without filters with the difference being stemming of ‘universities’ in MarkLogic. |
| AFFIL4 | 86 | 666 | affil (university of york). Likely attributed to differences in publication counts as well as identification of root nodes. MarkLogic and SOLR had similar counts (1087/1075) without the filter with the difference being stemming of ‘universities’ in MarkLogic. |