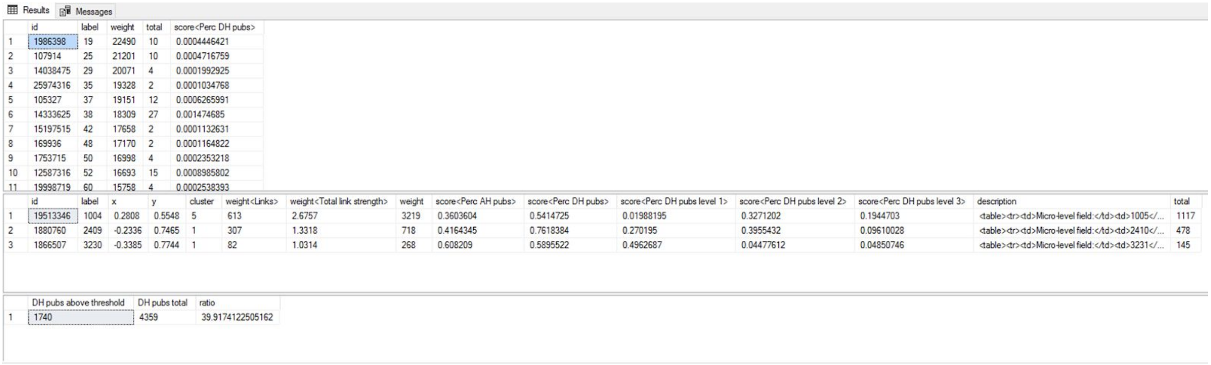
December 20th, 2019

For each cluster, if it has more than a certain percentage of DH publications, all of its publications are DH.

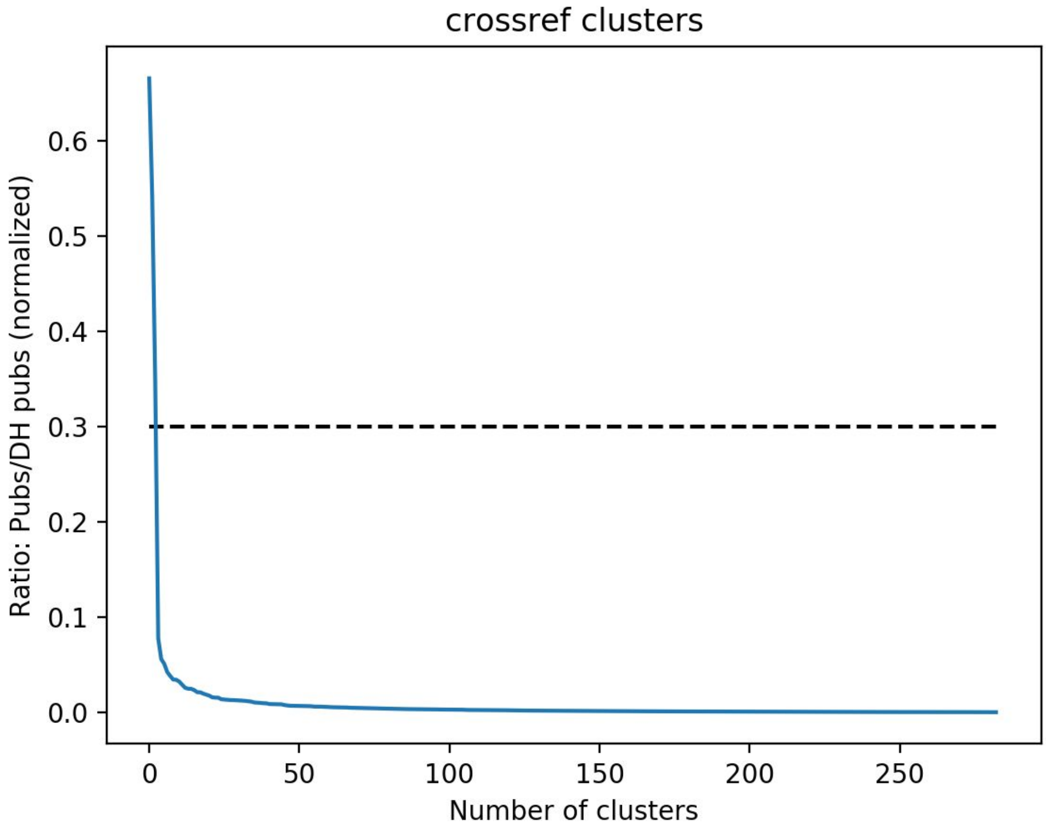
With crossref, we can have these results:
With a 30% threshold, we take into account 3 clusters and 40% of the DH publications from journals.



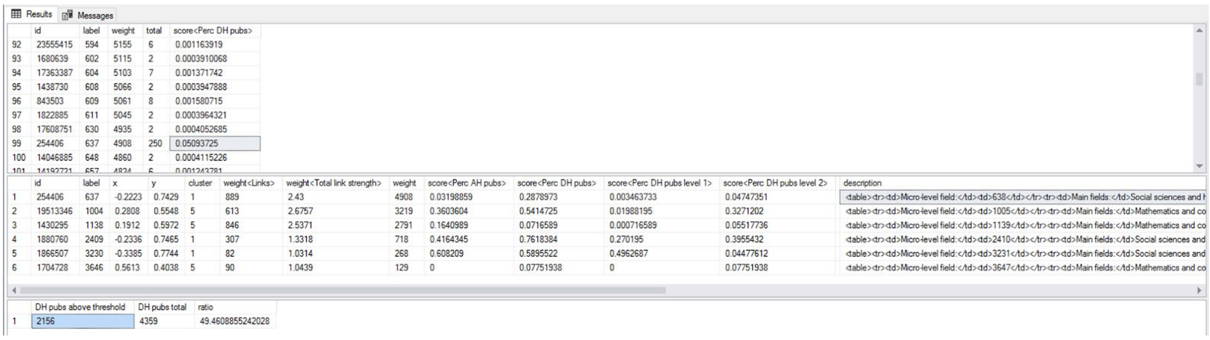
	id	label	weight	total	score<Perc DH pubs>
1	1986398	19	22490	10	0.0004446421
2	107914	25	21201	10	0.0004716759
3	14038475	29	20071	4	0.0001992925
4	25974316	35	19328	2	0.0001034768
5	105327	37	19151	12	0.0006265991
6	14333625	38	18309	27	0.001474685
7	15197515	42	17658	2	0.0001132631
8	169936	48	17170	2	0.0001164822
9	1753715	50	16998	4	0.0002353218
10	12587316	52	16693	15	0.0008985802
11	19998719	60	15758	4	0.0002538393

	id	label	x	y	cluster	weight<Links>	weight<Total link strength>	weight	score<Perc AH pubs>	score<Perc DH pubs>	score<Perc DH pubs level 1>	score<Perc DH pubs level 2>	score<Perc DH pubs level 3>	description	total
1	19513346	1004	0.2808	0.5548	5	613	2.6757	3219	0.3603604	0.5414725	0.01988195	0.3271202	0.1944703	<table><tr><td>Micro-level field:</td><td>1005</...	1117
2	1880760	2409	-0.2336	0.7465	1	307	1.3318	718	0.4164345	0.7618384	0.270195	0.3955432	0.09610028	<table><tr><td>Micro-level field:</td><td>2410</...	478
3	1866507	3230	-0.3385	0.7744	1	82	1.0314	268	0.608209	0.5895522	0.4962687	0.04477612	0.04850746	<table><tr><td>Micro-level field:</td><td>3231</...	145

	DH pubs above threshold	DH pubs total	ratio
1	1740	4359	39.9174122505162



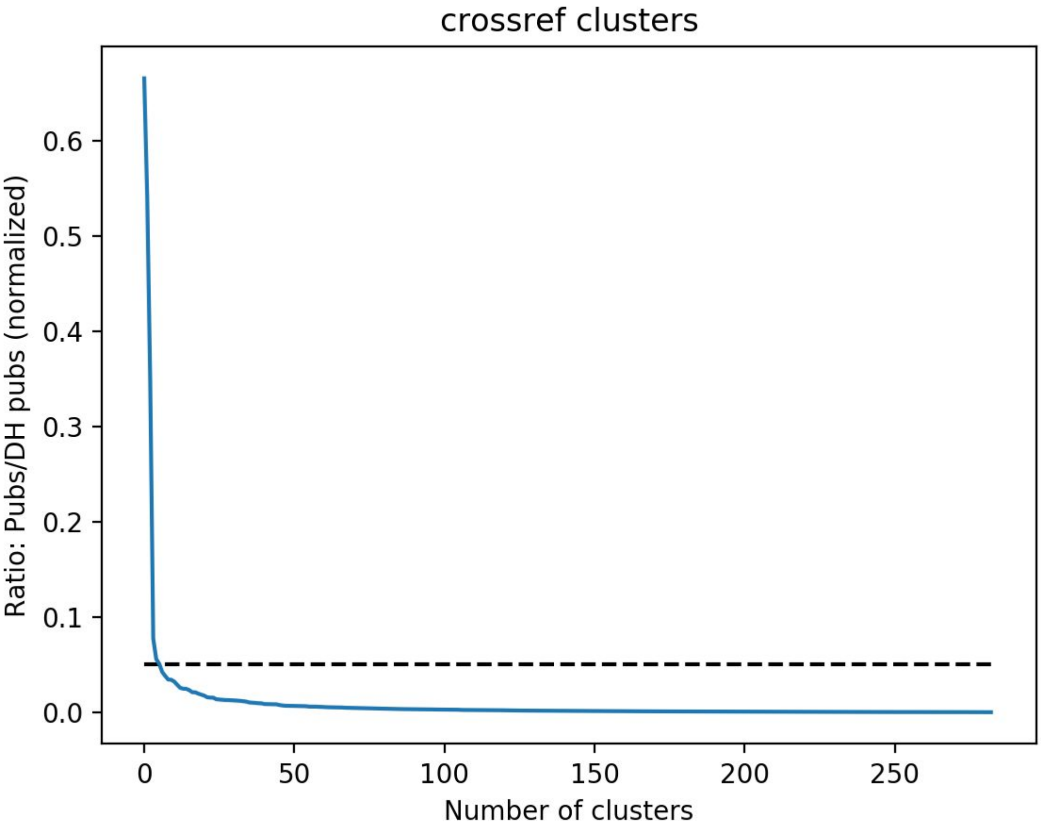
With a 5% threshold, we take into account 6 clusters and 49% of the DH publications from journals.



	id	label	weight	total	score<Perc DH pubs>
92	23555415	594	5155	6	0.001163919
93	1680639	602	5115	2	0.0003910068
94	17363387	604	5103	7	0.001371742
95	1438730	608	5066	2	0.0003947888
96	843503	609	5061	8	0.001580715
97	1822885	611	5045	2	0.0003964321
98	17608751	630	4935	2	0.0004052685
99	254406	637	4908	250	0.05093725
100	14046885	648	4860	2	0.0004115226
101	14192721	657	4824	6	0.001243781

	id	label	x	y	cluster	weight<Links>	weight<Total link strength>	weight	score<Perc AH pubs>	score<Perc DH pubs>	score<Perc DH pubs level 1>	score<Perc DH pubs level 2>	description
1	254406	637	-0.2223	0.7429	1	889	2.43	4908	0.03198859	0.2878973	0.003463733	0.04747351	<table><tr><td>Micro-level field:</td><td>638</td></tr><tr><td>Main fields:</td>Social sciences and h
2	19513346	1004	0.2808	0.5548	5	613	2.6757	3219	0.3603604	0.5414725	0.01988195	0.3271202	<table><tr><td>Micro-level field:</td><td>1005</td></tr><tr><td>Main fields:</td>Mathematics and co
3	1430295	1138	0.1912	0.5972	5	846	2.5371	2791	0.1640989	0.0716589	0.000716589	0.05517736	<table><tr><td>Micro-level field:</td><td>1139</td></tr><tr><td>Main fields:</td>Mathematics and co
4	1880760	2409	-0.2336	0.7465	1	307	1.3318	718	0.4164345	0.7618384	0.270195	0.3955432	<table><tr><td>Micro-level field:</td><td>2410</td></tr><tr><td>Main fields:</td>Social sciences and
5	1866507	3230	-0.3385	0.7744	1	82	1.0314	268	0.608209	0.5895522	0.4962687	0.04477612	<table><tr><td>Micro-level field:</td><td>3231</td></tr><tr><td>Main fields:</td>Social sciences and
6	1704728	3646	0.5613	0.4038	5	90	1.0439	129	0	0.07751938	0	0.07751938	<table><tr><td>Micro-level field:</td><td>3647</td></tr><tr><td>Main fields:</td>Mathematics and co

	DH pubs above threshold	DH pubs total	ratio
1	2156	4359	49.4608855242028



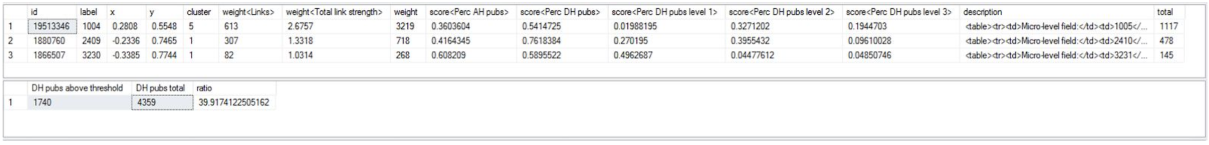
In this case, we have around 50% of sparse DH publications in fewer dense clusters.

Do we have to treat them in different ways?

Clusters (with 30+/5+ % DH pubs) are DH clusters, so all the pubs within are DH
DH sparse pubs, all the other pubs fewer dense. Analyze their cluster distribution and what they contain. Why these pubs are there.

Experiment 01

threshold	30%
DH pubs ratio	39.9% [1740 / 4359]
clusters	3



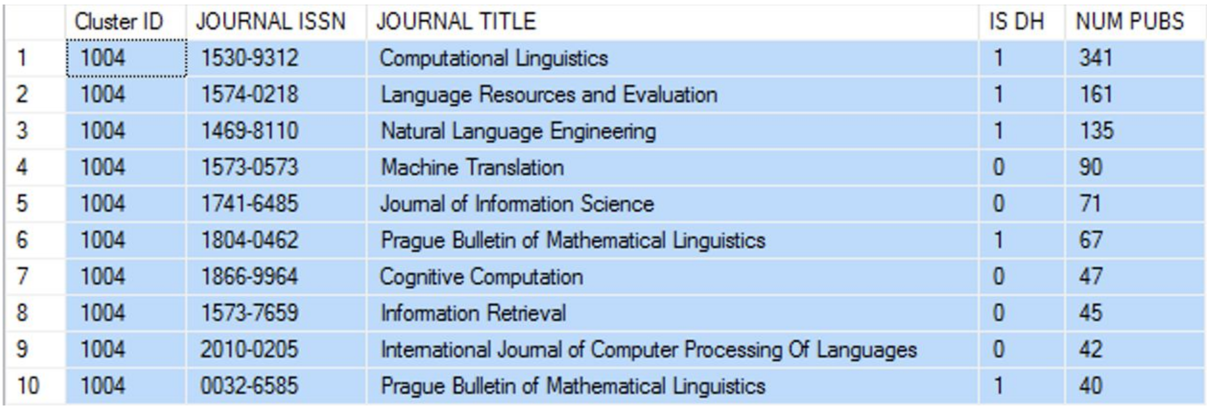
	id	label	x	y	cluster	weight<Links>	weight<Total link strength>	weight	score<Perc AH pubs>	score<Perc DH pubs>	score<Perc DH pubs level 1>	score<Perc DH pubs level 2>	score<Perc DH pubs level 3>	description	total
1	19513346	1004	0.2808	0.5548	5	613	2.6757	3219	0.3603604	0.5414725	0.01988195	0.3271202	0.1944703	<table><tr><td>Micro-level field:</td><td>1005</...	1117
2	1880760	2409	-0.2336	0.7465	1	307	1.3318	718	0.4164345	0.7618384	0.270195	0.3955432	0.09610028	<table><tr><td>Micro-level field:</td><td>2410</...	478
3	1866507	3230	-0.3385	0.7744	1	82	1.0314	268	0.608209	0.5895522	0.4962687	0.04477612	0.04850746	<table><tr><td>Micro-level field:</td><td>3231</...	145

	DH pubs above threshold	DH pubs total	ratio
1	1740	4359	39.9174122505162

Exploring deeper the clusters

Cluster 1

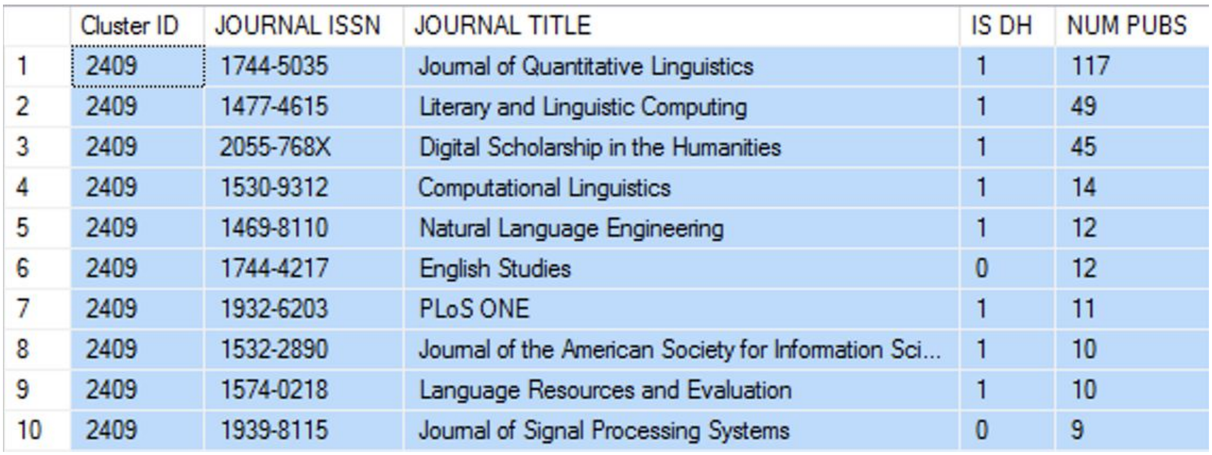
label: 1004
publications: 3219
journals: 536
Top 10 journals:



	Cluster ID	JOURNAL ISSN	JOURNAL TITLE	IS DH	NUM PUBS
1	1004	1530-9312	Computational Linguistics	1	341
2	1004	1574-0218	Language Resources and Evaluation	1	161
3	1004	1469-8110	Natural Language Engineering	1	135
4	1004	1573-0573	Machine Translation	0	90
5	1004	1741-6485	Journal of Information Science	0	71
6	1004	1804-0462	Prague Bulletin of Mathematical Linguistics	1	67
7	1004	1866-9964	Cognitive Computation	0	47
8	1004	1573-7659	Information Retrieval	0	45
9	1004	2010-0205	International Journal of Computer Processing Of Languages	0	42
10	1004	0032-6585	Prague Bulletin of Mathematical Linguistics	1	40

Cluster 2

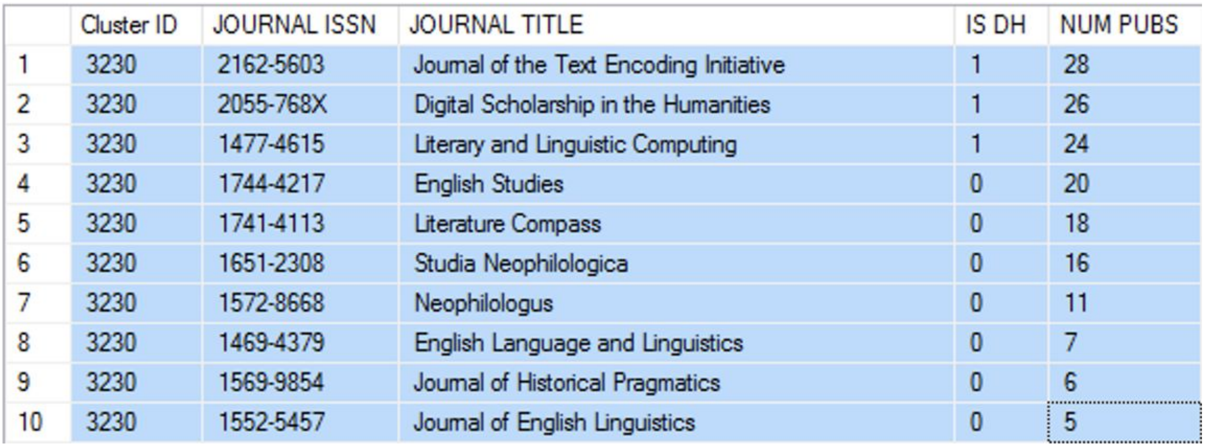
label: 2409
publications: 718
journals: 251
Top 10 journals:



	Cluster ID	JOURNAL ISSN	JOURNAL TITLE	IS DH	NUM PUBS
1	2409	1744-5035	Journal of Quantitative Linguistics	1	117
2	2409	1477-4615	Literary and Linguistic Computing	1	49
3	2409	2055-768X	Digital Scholarship in the Humanities	1	45
4	2409	1530-9312	Computational Linguistics	1	14
5	2409	1469-8110	Natural Language Engineering	1	12
6	2409	1744-4217	English Studies	0	12
7	2409	1932-6203	PLoS ONE	1	11
8	2409	1532-2890	Journal of the American Society for Information Sci...	1	10
9	2409	1574-0218	Language Resources and Evaluation	1	10
10	2409	1939-8115	Journal of Signal Processing Systems	0	9

Cluster 3

label: 3230
publications: 268
journals: 82
Top 10 journals:



	Cluster ID	JOURNAL ISSN	JOURNAL TITLE	IS DH	NUM PUBS
1	3230	2162-5603	Journal of the Text Encoding Initiative	1	28
2	3230	2055-768X	Digital Scholarship in the Humanities	1	26
3	3230	1477-4615	Literary and Linguistic Computing	1	24
4	3230	1744-4217	English Studies	0	20
5	3230	1741-4113	Literature Compass	0	18
6	3230	1651-2308	Studia Neophilologica	0	16
7	3230	1572-8668	Neophilologus	0	11
8	3230	1469-4379	English Language and Linguistics	0	7
9	3230	1569-9854	Journal of Historical Pragmatics	0	6
10	3230	1552-5457	Journal of English Linguistics	0	5