UNIVERSITY OF TORONTO

MUNK SCHOOL OF GLOBAL AFFAIRS

Join the Global Conversation

The Citizen Lab

Research Brief
August 2013

Who's the Boss? The Difficulities of Identifying Censorship in an Environment with Distributed Oversight: A Large-Scale Comparison of Wikipedia China with Hudong and Baidu Baike

Author: Jason Q. Ng

INTRODUCTION

In 2008, Baidu's chief scientist William Chang said, "There's, in fact, no reason for China to use Wikipedia . . . It's very natural for China to make its own products." Today Hudong (baike.com) and Baidu Baike (baike.baidu.com) greatly eclipse the Chinese-language version of Wikipedia despite (or because of) the censorship known to take place on the sites. However, identifying outright instances or patterns in censorship can be difficult due to the (mostly) user-generated nature and oversight of the content. Instead, this project attempts to perform a large-scale comparison of the three services, matching thousands of Chinese-language Wikipedia articles with their in-China counterparts, in order to identify the "content gaps" in the two *baike* (Chinese for "encyclopedia," which we use to refer to Hudong's and Baidu's online encyclopedias). Censorship—or at the very least anomalies in the generation of content—*might* be identified by articles that don't exist, "protected" articles that are not editable by regular users, and by articles that are much shorter than those on Wikipedia China. The reason *might* is emphasized is due to the distributed oversight nature of these online encyclopedias, where not only governments but also companies and users get to play the role of content gatekeeper. This decentralization makes attributing who is responsible for apparent censorship more difficult, a topic which this report will explore in detail by examining how it functions in these online encyclopedias.

In addition to the exploring the difficulties in identifying censorship, this post will also lay out the research methodology of the article matching, some of the initial results, and the next steps to be taken in the coming months as we continue to analyze this data and expand the project. As the data for this project has just been collected, this report is more data dump than fully-composed, critical analysis (see the final section for future avenues of research and thinking), but hopefully it serves as an introduction into the sorts of quality data available in this project.

Tables with lists of articles that are protected/locked on Wikipedia, Baidu Baike, and Hudong as well as a list of articles that are found on Wikipedia China but not found on the two *baike* are below. You can jump directly to them by clicking the links in the previous sentence, but as the data is still preliminary and has many limitations, one might be best served reading through the following sections.

BAIKE: CHINESE ENCYCLOPEDIAS

As William Chang of Baidu foretold, mainland Chinese netizens have gravitated toward local products such as Hudong and Baidu Baike, leaving Wikipedia China to be edited and read primarily by users in Taiwan, Hong Kong, and the rest of the Chinese diaspora. Today, in terms of raw visitors and article count, Hudong and Baidu Baike dwarf Wikipedia China, which has roughly 700,000 articles versus over 5 million in each of the two *baike*. Certainly, while China's sporadic blocking of access to Wikipedia at various points over the past ten years has certainly been a factor in limiting Wikipedia China's growth among mainland users, Baidu and Hudong's dominance may be more credited to Baidu's entrenched position as the dominant search engine in China (thus allowing for cross-site "partnerships" and synergies[1]) and Hudong's bevy of features built into its custom wiki and social networking platform.

Though the *baike* are incredible sources of information on China, they have been dogged by allegations that they have liberally "borrowed" content from other websites, including Wikipedia–a not damnable offense in and of itself since Wikipedia's content is free to share and re-use, but the two *baike* are for-profit and inform users that any content contributed is property of Hudong and Baidu. In the past, Baidu was noted as the worst offender, plagiarizing from not only Wikipedia without credit but also from Hudong.[2] Though the goal of this project is not to analyze these plagiarization claims and to quantify the amount of shared content among the three encyclopedias, the data generated from the matching of articles as explained in the methodology section below would allow someone to easily perform this sort of follow-up analysis once the data from this project is cleaned up and released.

THE DIFFICULTIES OF IDENTIFYING CENSORSHIP IN AN ENVIRONMENT WITH DISTRIBUTED OVERSIGHT

This project began with a question: everyone "knows" Hudong and Baidu Baike, like all Chinese websites, have to restrict certain kinds of content on their websites, but is it possible to empirically prove that censorship is taking place on the sites? Thinking about different ways of testing our assumption with the publicly available data on the websites drove this project.

First, what would we consider signs of censorship on Hudong and Baidu Baike? The most obvious would be the lack of certain articles on topics that are known to be notable. Thus, using Wikipedia China as a control, we can propose that if an article on say 上海帮 (The Shanghai Gang) exists on Wikipedia China, barring censorship, it should exist on Hudong and Baidu Baike, especially since Hudong and Baidu Baike have a much larger library of entries. However, attributing the lack of an entry due to censorship is not a perfect science since Wikipedia China itself isn't a perfect control–though entries in Wikipedia China are assumed to be of interest to Hudong and Baidu Baike users, and thus should have articles in those encyclopedias, one should keep in mind that Wikipedia China does tend to have a Taiwanese bend due to its userbase. However, using missing articles as a potential indicator for possible censorship–especially if the article that is missing is a long one–is a reasonable start.

Second, comparisons could also be made between the length of the articles between the encyclopedias. For instance, an article might exist on all three services, but they might be drastically shorter than their Wikipedia counterpart. For instance, the main body of the Wikipedia entry for 艾未未 (Ai Weiwei) is over 20,000 characters long (spaces removed) while the Baidu entry for him clocks in at 2,000 characters and the Hudong one at 3,500–this is despite the fact that article lengths for the Baidu articles sampled in this project are on the whole longer than Wikipedia's: the mean Wikipedia China article length is 8,174 characters (median: 4,207) while Baidu's is 12,679 (median: 9,029). Discrepancies of the sort in the Ai Weiwei article might simply be a case of greater interest in the topic outside mainland China than within, or, again, it might be another potential

indicator of censorship.

Third, some especially sensitive or controversial articles are unable to be edited except by users with much higher privileges than ordinary members. Being unable to change an article is not in and of itself a sign of censorship; for instance, Wikipedia "protects" certain articles to prevent vandalism and pointless back-and-forth "edit wars." Baidu and Hudong no doubt have similar intentions in mind as well, but what matters here is matter of transparency–while Wikipedia publishes a list of all protected pages, as far as we can tell, no such corresponding list exists for Hudong and Baidu Baike–and authority. Was it the choice of editors and users at Hudong and Baidu to classify certain articles as locked or was the decision made higher? Was there an open discussion of such matters or was a list handed down from somewhere above? Interviewing regular users of Baidu Baike and Hudong might provide insight into such questions, but for now, we have some data to start with.

Finally, there are more subtle ways to disrupt access to information, many of which we will no doubt uncover as we continue to sift through the data. One that we've noticed is the "failure" by Hudong and Baidu to redirect certain article titles the same way that Wikipedia does. For instance, a search for 艾神, a laudatory nickname meaning "God Ai" for Ai Weiwei, properly redirects to the Wikipedia article for him. Hudong and Baidu Baike don't perform such re-directs. Again, whether this is a conscious decision or merely an inadvertent one cannot be answered by looking at this one example. However, by looking at such cases in the aggregate one might be able to make a more legitimate claim that something might be going on.

Many new media outlets such as Sina Weibo privilege users with the ability to generate the content that goes on the website–in essence, to be not only their own programmer or broadcaster but also the producer. However, because such websites host their users' content, they are also in charge of regulating and ensuring that such content complies with all Chinese laws–regardless of how vague such regulations might be. Thus attributing censorship that takes place on these sites can be unclear–is it the government that mandated certain topics are off-limits or is it the company that restricts the content?–an intentional feature of the decentralized system of information control that Chinese authorities have developed.

Censorship is further distributed on the *baike* because now not only are users their own programmer and producer, but they also serve in an oversight capacity as an editor. Unlike Wikipedia, users who aren't registered can't begin editing and creating articles, but for the most part, registered users can edit most general articles, and as they engage with the site longer, they achieve greater and greater levels of ability to edit and oversee the website. Thus, there could are always at least three potential reasons for why an article doesn't exist, an article is shorter, or an article is locked on Hudong or Baidu Baike: government entities, private companies, or users themselves. Judging whether or not these factors are genuine instances of governmental censorship or due to explainable, organic reasons can be quite tricky. Because of the multiple layers of oversight, what may appear to be outright censorship may be a less malicious (though no less pernicious) case of self-censorship.

METHODOLOGY

As mentioned, this project is a large-scale attempt to quickly and automatically match thousands of Wikipedia China articles with their corresponding Hudong and Baidu Baike entry. A script was developed which did just that, taking a keyword, locating the correct Wikipedia article and scraping the desired data, and then repeating the process for Hudong and Baidu, respectively, before moving on to the next keyword–essentially relying on Hudong and Baidu to perform the proper title matching or, if the title did not match exactly, redirect to the appropriate article.

Zhichun Wang, Juanzi Li, Zhigang Wang, and others at the Computer Science Department at Tsinghua

University have already made great strides in "knowledge linking" between English Wikipedia and Hudong/Baidu Baike, and their approach to matching articles across different languages using machine learning to read semantic information is well beyond the scope of this report. However, techniques from their work may be employed in the future, though the current title matching approach employed for this projecy is already a fairly robust solution as is–especially since we are dealing with only one language. In "Cross-lingual knowledge linking across wiki knowledge bases," Wang et al. reported that simple title matching gave them a precision percentage in excess of 99% even when having to translate from Chinese to English; however the recall rate was an atrocious 32%.

The data collected for this project still has to be evaluated properly, but thus far results seem very promising. Out of 5,143 keywords tested, Hudong and and Baidu each successfully found or redirected us to an appropriate article 3,539 and 3,600 times respectively, a recall rate of at least 68%–a number which is likely much higher once we account for the fact that the script also attempted to use search results to identify and match instances when the keyword was found in the returned search results' snippets (the article titles and summaries).

Of these 3,500+ times, the exact title–meaning a character for character match–was made between Wikipedia and Hudong 2,037 times, and 2,080 times between Wikipeda and Baidu Baike. If one assumes that an exact match in the article title is a match in the article content as well, then the precision rating for matched articles is nearly 60% already, and that's not yet taking into account all the articles titles which are only slightly different between the services, either due to different styling conventions (e.g. Wikipedia's article on national universities is titled with traditional characters [國立大學] while Hudong and Baidu both use simplified [国立大学]) or other minor variations. Overall, the matching procedure, at least for the terms tested thus far, seems more than acceptable for the goals of this project, though of course attempts will be made to verify the precision and increase the recall rate.

As a test of the script and the methodology, a sample of 5,143 keywords involving a mix of topics known to be sensitive and those known to be popular were generated from five different sources: the titles of articles protected by Wikipedia (282); article titles from the list of Wikipedia China articles on GreatFire.org's watchlist of censored Wikipedia pages (489); article titles of the top 1000 most viewed articles on Wikipedia China in April 2013 (995 after a few Wikipedia meta-pages were removed); the titles of articles that generated more than a total of 10 combined views on August 1, 0:00-1:00 and 12:00-13:00 (3,470); and finally, the only non-Wikipedia source, a list of keywords from the website Blocked on Weibo previously confirmed to have been prevented from returning search results on Sina Weibo (840). Some of the final sample of keywords appeared on multiple lists, and the overlap of four of the sources is shown below (the source of words from August 1 were dropped because 5-way Venn diagrams are not as pretty…).

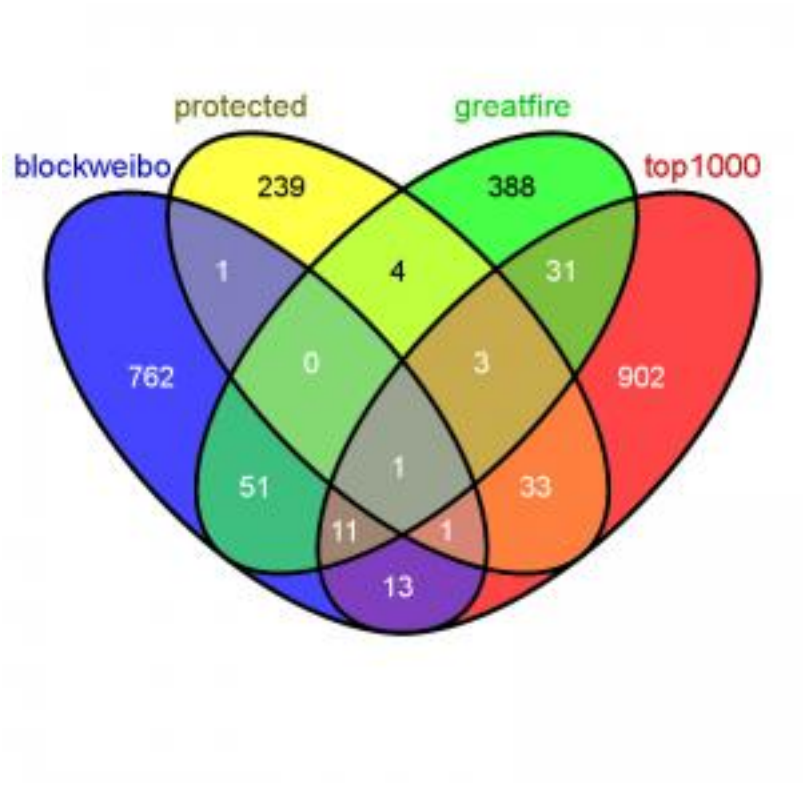

Figure 1: 4-way Venn diagram showing overlap of sources used for keywords (not shown: keywords taken from most viewed Wikipedia China articles on Aug 1)

PROTECTED, MISSING, AND CENSORED(?) CONTENT

As mentioned previously, an article that is protected or locked doesn't necessarily mean that it is being censored–Wikipedia often uses article protection as a means to prevent malicious behavior. However, the locking down of articles itself can be abused and used in a malicious manner, especially if the decision to do so is non-transparent, arbitrary, and/or doesn't reflect the sentiment of the group. In such cases, Wikipedia fortunately has active "talk pages" which allow all users to debate and contest such issues. Baidu Baike also has discussion pages which allow users to to something similar: for instance, here's one for the entry on the United Kingdom wherein users quibble about whether or not England can lay claim to being the place where the first "bourgeois revolution" took place (the consensus was no, the Dutch Revolt came before it). However, many protected Baidu Baike entries contain no such talk pages. For instance, trying to reach Xi Jinping's talk page returns an error message. Hudong eschews separate talk pages in favor of comments at the bottom of each entry. However, again, oftentimes for sensitive and/or protected entries, there is no opportunity to leave your comment; the comments box simply doesn't appear.

Furthermore, while the list of pages protected by Wikipedia is published by the site, as far as we can tell, no such list exists for Hudong or Baidu Baike. Thus, even with he caveats noted above about how the articles in the following three tables don't necessarily conclude that censorship is taking place in those entries, this exercise still seems like a useful one if only to make such information publicly available. Again, proper analysis is yet to be done on what is protected/locked and what sorts of patterns exist in the three different encyclopedias approaches to article protection.

Note: the machine translations come from Google Translate. Those that have been corrected or verified have notes or a dot in the third column; if the third column is empty, the translation is still to be confirmed–an ongoing part of this project. Please do not disseminate unconfirmed machine translations without first verifying that they are correct.

Protected articles on Wikipedia China

Article title + Wikipedia link	Machine translation	Human translation / notes (if field is blank, translation still to be confirmed)
民主黨_(香港)	Democratic Party (Hong Kong)	.
庾澄慶	Harlem Yu	Taiwanese pop star
東方報業集團	OrientalPress Group	Hong Kong publisher of newspapers: Oriental Daily News
丁部領	Dinh Bo Linh	Vietnamese emperor
蒼井空	Sora Aoi	Japanse porn star
蔡煌瑯	Tsai Huang-lang	Taiwanese politician
首页	homepage	.
蔡英文	Tsai Ing-wen	
新店救護車阻擋事件	Storeambulance blocking event	
草榴社区	Grass garnet Community	Caoliu online forum
丁文雄	DingwenXiong	
AV女優	AV Actress	.
南陽郡	NanyangCounty	
林書豪	Jeremy Lin	basketball player
阮文雄	RuanwenXiong	
星野亞希	Aki Hoshino	
濱田翔子_(演員)	Bin Shoko(Actor)	
佐佐木希	Sasaki Nozomi	

朝河兰	NorthRiver Portland	
濱崎步	Ayumi Hamasaki	Japanese porn star
原紗央莉	OriginalSaori	
小澤瑪麗亞	Maria Ozawa	
Angelababy	Angelababy	
佐山愛	Ai Sayama	
郭書瑤	GuoShu-Yao	
大纪元时报	The Epoch Times	Falun Gong-connected publication
國立大學	national university	.
公立大學	Public universities	.
全球定位系统	GlobalPositioning System	
国父_(罗马帝国)	Pater Patriae (Roman Empire)	“Father of the Country” honorific
潘佩珠	Phan	
哥德巴赫猜想	Goldbach’s conjecture	
彭淮南	Perng	
吳淑珍	Wu Shu-chen	
第十四世达赖喇嘛	FourteenthDalai Lama	
宋教仁	Sung	
司马南	Sima Nan	
天海翼	Day sea wing	
性高潮	Orgasm	
蔡依林	Jolin	
满族	Manchu	
U-KISS	U-KISS	
维基百科	Wikipedia	

印度神油	Indian god oil	
江泽民	Jiang	
东京热	Tokyo Hot	
正覺同修會	ChingPractitioners Association	
六合彩	LOTTO	
越南政黨列表	Vietnameseparty list	
干	Dry	
朱明	Zhu	
中华人民共和国行政区划	People's Republic of administrativedivisions	
齡記書店	Ling Kee	
2012年香港立法會選舉	2012 Hong Kong Legislative CouncilElection	
真佛宗	TrueBuddha School	
海椰子	Sea coconut	
鶴佬陸上扒龍船	GrilledHoklo dragon boat onshore	
馮光遠	Peng says	
蔡瑞月	Tsai	
衛星定位系統	Satellite positioning system	
大连理工大学	DalianUniversity of Technology	
拌麵	Noodles	
MediaWiki	MediaWiki	
利菁	Lee Ching	
狭义相对论	SpecialRelativity	
南京大學校友列表	Nanjing University Alumni List	
李长春	LiChangchun	
有栖川宮	Arisugawanomiya	
郑州加州工业城	ZhengzhouCity of Industry, California	

莫莉花	Molly Flowers	
曾偉恩	Zeng Wayne	
捕捉、絕育、釋放	Trap-Neuter-Release	
自慰	Masturbation	
顺阳范氏	Shun Yang Fan	
港西镇_(崇明县)	Hong KongWest Town (Chongming County)	
陳炳	Chen Ping	
發正念	HairMindfulness	
許瑜真	Xu Yu True	
十二年國教學生研討會	Twelvenational seminar teaching students	
任天堂溥天	Nintendo Popteam	
南阳范氏	NanyangFan	
阿坎巴羅雕像	Aquin Barrow Statue	
伊卡黑石	IcaBlackstone	
痘弦	Ya Xian	
呂泉生	LvQuansheng	
中国大陆	Chinese mainland	
丁先皇	DingXianhuang	
天主教香港教區	Catholic Diocese of Hong Kong	
蔡衍明	Tsai	
小游戏	Small game	
曹昂	Cao Ang	
曹鑠	Cao Shuo	
加拉帕戈斯象龜	Galapagostortoises	
北西摩島	North Seymour Island	
Hao123网址之家	Hao123website	

中华人民共和国政府认定的邪教组织列表	PRC Government has identified alist of cult	
被政府認定為邪教的團體列表	Identifiedby the government as a cult group list	
馬三家女子勞教所	Masanjia Women's Forced Labor Camp	
神韵艺术团	Shen YunPerforming Arts	
多面体	Polyhedron	
巴丹群島	BatanesIslands	
蘇家屯事件	Sujiatun	
薄瓜瓜	Bo Guagua	
拳王_(電視劇)	Muhammad (TV series)	
朱雪璋	ZhuxueZhang	
黑魔女學園	Black witch academy	
缺宅男女	Lack ofhouse men and women	
楊怡	Tavia	
伏見宮	FushimiPalace	
对毛泽东的评价	Evaluation of Mao	
中国共产党中央委员会主席毛泽东同志支持美国黑人抗暴斗争的声明	ChineseCommunist Party Central Committee with Comrade Mao Zedong declared thestruggle to support African-American uprising	
蟾蜍	Toad	
闲院宫载仁亲王	CourtPalace, Prince Akishino idle load	
皇室	Imperial family	
當旺爸爸	When thebusy dad	
李克强	Li Keqiang	
幾米	Jimmy	
维权运动	Rights movement	

骨部	Bony part	
行政院環境保護署	Environmental Protection Agency	
南華足球隊	SouthChina Football Team	
鍾嘉欣	Linda	
行政院	ExecutiveYuan	
朝鲜战争	Korean War	
蔡宇傑	Cai Yujie	
越南共和国	Republic of Vietnam	
陳智雄	ChenZhixiong	
陈良宇	Chen Liangyu	
單戀雙城	Unrequitedlove Twins	
交通部中央氣象局	Central Weather Bureau	
中華民國教育部	Republicof China Ministry of Education	
中華民國交通部	Republic of China Ministry ofTransportation	
巨輪	Large ship	
九江十二坊	Jiujiang twelve Square	
香港獨立媒體	HongKong’s Independent Media	
方志敏	Fang Zhimin	
梁振英	LeungChun-ying	
表情符号	Emoticons	
雷霆掃毒	Thunderantidrug	
摘星之旅	Reaching for the Stars Tour	
Yummy_Yummy	YummyYummy	
宣萱	Jessica	
賭博	Gambling	
澳門博彩業	Macau gaming	

台灣日治時期	ColonialTaiwan	
星光少女_Rainbow_Live	Starlight Girls Rainbow Live	
葉瑋庭	Yewei Ting	
李若彤	Carman	
宜昌市	YichangCity	
萬曆朝鮮之役	Wanli Korean Battle	
露梁海戰	BeamedBattle	
台北暗殺星	Taipei assassination Star	
閑院宮	Busyhospital Palace	
时空轮回	Space reincarnation	
叛逃_(電視劇)	Defection(TV series)	
樊光耀	Fan Guangyao	
忠奸人	DonnieBrasco	
寒山潛龍	Hanshan Qianlong	
胡定欣	Nancy	
在台越南人	Vietnamese people in Taiwan	
洪仲丘事件	Hung ChungConfucius event	
王亦豐	Wang Yifeng	
臺北市私立再興高級中學	PrivateHigh School, Taipei redevelop	
吳奇隆	Nicky	
鄭嘉穎	Kevin	
朴善英	Sun Young	
崔真理	CUI Truth	
翡翠台電視劇集列表_(2010年代)	Jade TV set list (2010s)	
夏侯霸	Xiahou Ba	

大韓帝國	Daehan Empire	
李朝實錄	Chao-Record	
吳卓羲	Ron Ng	
東華三院邱子田紀念中學	TWGH YauTze Tin Memorial College	
國立暨南國際大學	National Chi Nan University	
渤海国	BohaiState	
神探高倫布	Detective high Columbus	
李宇春	Li Yuchun	
余貴美子	Kimiko	
Running_Man	RunningMan	
小高句丽	Small Koguryo	
安东都督府	AntonDudufu	
唐与高句丽的战争	Tang Dynasty and Koguryo war	
EXO	EXO	
馬賽_(藝人)	Marseille (entertainer)	
天下女人心	WorldWomen Want	
东淤地站	East warping Station	
上海市	Shanghai	
半总统制	Semi-presidentialism	
銀正雄	SilverMasao	
倪妮	Ni Ni	
十四世达赖	Dalai	
李長春	Li Changchun	
民主黨_(香港)	Democrats_ (Hong Kong)	
江澤民	Jiang	
苏家屯事件	Sujiatun	

薄熙来	Bo Xilai	
香港民主黨	Hong KongDemocratic Party	
丹增嘉措	Tenzin Gyatso	
Wikipedia	Wikipedia	
Hao123	Hao123	
日本AV女優	JapaneseAV Actress	.
鐵馬尋橋	A Fistful Of Stances	
東京熱	Tokyo Hot	
蘭陵王_(電視劇)	Lanling (TV series)	
洪仲丘	Hung ChungConfucius	
薄熙来事件	Bo Xilai event	
維基百科	Wikipedia	
手淫	Masturbation	
溏心風暴	Heart ofGreed	
科比·布莱恩特	Kobe Bryant	
SUHO#SUHO	SUHO	
貓屎媽媽	Mother cat feces	
守業者	Entrepreneurs	
TPA	TPA	
父與子_(民視電視劇)	Father andSon (FTV TV series)	
武藤兰	Muto blue	

Protected articles on Baidu

Article title + Baidu Baike link	Machine translation	Human translation / notes (if field is blank, translation still to be confirmed)
丹增嘉措	Tenzin Gyatso	14th Dalai Lama
司马南	Sima Nan	Television pundit
伟哥	Viagra	
江泽民	Jiang	
六合彩	LOTTO	
李长春	Li Changchun	
自慰	Masturbation	
薄瓜瓜	Bo Guagua	
李克强	Li Keqiang	
部骨陷	Ministry bone defect	
陈良宇	Chen Liangyu	
中华民国教育部	Republic of China Ministry of Education	
梁振英	Leung Chun-ying	
叶玮庭	Yewei Ting	
宜昌	Yichang	
吴奇隆	Nicky	
李宇春	Li Yuchun	
万里	Miles	
三年自然灾害	Three years of natural disasters	
习仲勋	Xi	
习近平	Xi Jinping	

令计划	Orders Scheme	
何俊仁	Albert	
余杰	Yu Jie	
侯德健	Hou Dejian	
侯德健	Hou Dejian	
俞正声	Yu	
刘云山	Liu Yunshan	
刘延东	Liu Yandong	
刘慧卿	Emily	
刘淇	Liu Qi	
刘晓波	Liu Xiaobo	
丹增嘉措	Tenzin Gyatso	
反华势力	Anti-China forces	
吴邦国	Wu Bangguo	
周永康	Zhou Yongkang	
唯色	Woeser	
回良玉	Hui Liangyu	
国际特赦组织	Amnesty International	
宋任穷	Song Renqiong	
宋任穷	Song Renqiong	
宋彬彬	Song Binbin	
张德江	Zhang	
张高丽	Zhang Gaoli	
徐才厚	Xu Caihou	
李卓人	Yan	
李源潮	Li Yuanchao	

李瑞环	Li	
李瑞环	Li	
李长春	Li Changchun	
李鹏	Li Peng	
江泽民	Jiang	
汪洋	Wang Yang	
温家宝	Wen Jiabao	
激流中国	Torrent China	
热比娅·卡德尔	Rebiya Kadeer	
热比娅·卡德尔	Rebiya Kadeer	
王乐泉	Mr Wang	
王兆国	Wang	
王岐山	Wang Qishan	
王有才	Wang Youcai	
王乐泉	Mr Wang	
王立军	Wang Lijun	
秦城监狱	Qincheng Prison	
章嘉呼图克图	Akiyoshi Hutuketu	
章诒和	Zhang Yihe	
绿坝-花季护航	Green Dam – Youth Escort	
罗干	Luo Gan	
罗干	Luo Gan	
习近平	Xi Jinping	
胡耀邦	Hu Yaobang	
胡锦涛	Hu	
艾未未	Ai Weiwei	

华国锋	Hua Guofeng	
蒋介石	Chiang Kai-shek	
薄熙来	Bo Xilai	
西藏	Tibet	
贾庆林	Jia Qinglin	
贺国强	He Guoqiang	
贾庆林	Jia Qinglin	
赵紫阳	Zhao Ziyang	
赵紫阳	Zhao Ziyang	
达赖喇嘛	Dalai Lama	
郭伯雄	Guo	
阎明复	Yan Mingfu	
霍英东	Henry Fok	
马英九	Ma Ying-jeou	
丹增嘉措	Tenzin Gyatso	
阿沛·阿旺晋美	Ngapoi Ngawang Jigme	
中华人民共和国宪法	Constitution of the PRC	
刘慧卿	Emily	
李鹏	Li Peng	
张立昌	Zhang Lichang	
9·21乌坎村事件	9.21 Wukan event	
中国共产党	Communist Party of China	
绿坝-花季护航	Green Dam – Youth Escort	
彭丽媛	Peng Liyuan	
章诒和	Zhang Yihe	
红色高棉	Khmer Rouge	

中华民国	Republic of China	
中国共产党中央政治局	Chinese Communist Party Politburo	
中文维基百科	Chinese Wikipedia	
中华民国	Republic of China	
何清涟	He Qinglian	
共产党	Communist party	
华国锋	Hua Guofeng	
大跃进	Great Leap Forward	
天与地	Heaven and Earth	
林昭	Lin Zhao	
毛泽东	Mao Zedong	
民主	Democracy	
邓小平	Deng Xiaoping	
64	64	
tor	tor	
百度	Baidu	
百度	Baidu	
禽流感	Avian Influenza	
qq	qq	
泰国	Thailand	
文化大革命	Cultural Revolution	
母亲节	Mother's Day	
越南	Vietnam	
新加坡	Singapore	
阴茎	Penis	
H7N9型禽流感	H7N9 avian influenza	

尖锐湿疣	Genital warts	
周恩来	Zhou Enlai	
林彪	Lin Biao	
中国好声音	China good voice	
蒋介石	Chiang Kai-shek	
金正日	Kim Jong Il	
qq	qq	
铊中毒	Thallium poisoning	
谷俊山	Dragon Hill Valley	
阴蒂	Clitoris	
qq	qq	
吴千语	Wu thousand words	
金泫雅	Kim HyunA	
蒋经国	Chiang Ching-kuo	
铊	Thallium	
刘志军	Liu Zhijun	
射精	Ejaculation	
Nichkhun Horvejkul	Nichkhun Horvejkul	
元朝	Yuan	
潮吹	Squirting	
李小龙	Bruce Lee	
乳房	Breast	
朴施厚	Park Shi Hoo	
江青	Jiang Qing	
伊斯兰教	Islam	
毛新宇	Mao Xinyu	

张志新	Zhang Zhixin	
叶剑英	Ye Jianying	
亚斯伯格症候群	Asperger's Syndrome	
阿尔茨海默病	Alzheimer's disease	
中国人民解放军空军	People's Liberation Army Air Force	
处女膜	Hymen	
带状疱疹	Shingles	
自闭症	Autism	
斯里兰卡	Sri Lanka	
梅毒	Syphilis	
四川	Sichuan Province	
刘少奇	Liu Shaoqi	
胡春华	Hu Chunhua	
邓力群	Deng Liqun	
F-22战斗机	F-22 fighter	
乌有之乡	Utopia	
包皮	Foreskin	
汉朝	Han	
唐嫣	Tang Yan	
中国国民党	Chinese Kuomintang	
朱镕基	Zhu	
谭咏麟	Alan Tam	
狂犬病	Rabies	
飞蚊症	Floaters	
蔡英挺	Cai handsome	
彭德怀	Peng	

中暑	Heatstroke	
精液	Semen	
林允儿	Lin Yun children	
花木兰	Mulan	
青光眼	Glaucoma	
中国好声音第二季	China good sound second quarter	
吴昌德	Wu Changde	
三级片	Three pieces	
亨丁顿舞蹈症	Huntington	
尚福林	Shang Fulin	
黄华华	Huang Hua	
纳尔逊·罗利赫拉赫拉·曼德拉	Raleigh, Hela He pulled Nelson Mandela	
泌尿道感染	Urinary tract infections	
台北	Taipei	
达姆弹	Dumdum	
隐翅虫	Paederus	
痛风	Gout	
中华民国国军	Republic of China Armed Forces	
张又侠	Xia Zhang	
高虎城	Gao	
女性生殖系统	Female Reproductive System	
自慰	Masturbation	
红斑性狼疮	Lupus Erythematosus	
谵妄	Delirium	
百度百科	Baidu Encyclopedia	

湿疹	Eczema	
火药	Gunpowder	
糖尿病	Diabetes	
马馱	Ma Wen	
乌克兰	Ukraine	.
幽门螺杆菌	Helicobacter pylori	bacteria in stomach
杨尚昆	Yang	
贴吧	Post Bar	
艾滋病	AIDS	
乳糖不耐受	Lactose intolerance	
党和国家领导人	Party and state leaders	
前列腺	Prostate	.
北海舰队	North Sea Fleet	
月经	Menstruation	
李先念	Li Xiannian	Chinese politician, former president
类风湿关节炎	Rheumatoid Arthritis	
中国人民解放军第38集团军	38th Chinese People's Liberation Army	
乳癌	Breast cancer	.
低血压病	Hypotension	
刘亚洲	Liu Yazhou	
大麻	Marijuana	
火龙果	Pitaya	
王立军	Wang Lijun	
西藏	Tibet	
陈希同	Chen Xitong	.

三唑仑	Triazolam	drug for treating insomnia
便秘	Constipation	
强迫症	Obsessive-compulsive disorder	
耳鸣	Tinnitus	
双相障碍	Bipolar disorder	
阳痿病	Impotence	.
剑风传奇	Chuah legend	
狂犬病病毒	Rabies virus	
王震	Wang Zhen	Chinse politician, one of Eight Elders
精神分裂症	Schizophrenia	
苏州	Suzhou	
针眼	Pinprick	
韩寒	Han	
黑崎一护	Kurosaki Ichigo	
CF	CF	Crossfire, Korean video game
中国人民解放军总后勤部	People’s Liberation Army General Logistics Department	
公主病	Princess disease	
卵巢癌	Ovarian cancer	.
21-三体综合征	21 – trisomy	
子宫颈癌	Cervical cancer	.
尿酸	Uric acid	
房峰辉	Room Feng Hui	
牛肉	Beef	
肝炎	Hepatitis	

肺炎	Pneumonia	
胆固醇	Cholesterol	
血压	Blood pressure	.
黄金圣斗士	Gold Saint	
龟头	Glans	
AK-47突击步枪	AK-47 assault rifles	.
ps3	ps3	
TPA	TPA	
中华人民共和国副主席	Chinese Vice President	
中国共产党第一次全国代表大会	China First National Congress of the Communist Party	
中国共产党	Communist Party of China	
中华民国刑法	Republic of China Criminal Law	
凤凰卫视	Phoenix	
刘晓波	Liu Xiaobo	
前列腺按摩	Prostate massage	
大鲵	Giant salamander	.
孙政才	Sun Zhengcai	Chinese politician, head of Chongqing
彭真	Peng Zhen	
手足口病	HFMD	
拔火罐	Cupping	
日俄战争	Russo-Japanese War	
朱德	Zhu	
东森电视台	Eastern Television	
水肿	Edema	
爱奇艺	iQiyi	Chinese video

		website
番红花	Saffron	
疣	Wart	
痔	Hemorrhoid	
癌症	Cancer	
白内障	Cataract	
白血病	Leukemia	
睡眠	Sleeping	
睾丸	Testicle	
结核	Tuberculosis	.
花木兰	Mulan	
荨麻疹	Urticaria	form of hives (rash)
贾廷安	Jia Yanan	
邓朴方	Deng Pufang	son of Deng Xiaoping
高圆圆	Gao Yuanyuan	
中国共产党	Communist Party of China	
乔石	Qiao	
储波	Chu Bo	
冷溶	Leng Rong	
列确	Legqog	
吗啡	Morphine	
吴仪	Wu	
奸淫	Adultery	
宪章	Charter	
恋童	Pedophilia	
恋足	Foot Fetish	

逼	Force	
李群	Lie	
杜宪	DU Xian	
炸弹	Bomb	
游行	Parade	
丁羽心	Ding Yu heart	
万武义	Wan Wuyi	
伊斯兰教	Islam	
何厚铧	Ho	
何鲁丽	He Luli	
俞正声	Yu	
刘华清	Liu Huaqing	
刘振亚	Liu Zhenya	
中华人民共和国副主席	Chinese Vice President	
厉无畏	Li Wuwei	
叶选平	Ye Xuanping	
吴志明	Wu Zhiming	
吴阶平	Jieping	
夏德仁	Xia Deren	
姜异康	Jiang Yikang	
姜春云	Chunyun	
孟学农	Meng Xuenong	
全能神	Almighty God	
尉健行	Wei Jianxing	
张定发	Zhang Ding hair	
徐勤先	Xu Qin first	

戴相龙	Dai Xianglong	
自慰	Masturbation	
政治犯	Political	
暴露癖	Exhibitionism	
曾培炎	Zeng	
朱云来	Mr Zhu	
李肇星	Li Zhaoxing	
杜冷丁	Pethidine	
杜导正	Du Daozheng	
林嘉祥	Lin Jiaxiang	
梁保华	Liang Baohua	
TNT	TNT	
全能神	Almighty God	
中国共产党中央军事委员会	Chinese Communist Party Central Military Commission	
六月四日	04-Jun	
敏感	Sensitive	
台湾独立运动	Taiwan independence movement	
向巴平措	Qiangba	
奥克托今	Octogen	
幼幼新书	Youyou book	
江绵恒	Mr Jiang	
淫液	Cum	
滴蜡	Candle Wax	
焦国标	Jiao Guobiao	
王宝森	Wang Baosen	

由喜贵	Xigui	
白恩培	Bai Enpei	
巨型气枪	Huge Guns	
中国人民解放军第38军	38th Chinese People's Liberation Army	
经叔平	Jing Shuping	
色欲	Lust	
三硝基苯酚	Trinitrophenol	
董建华	Tung	
邓颖超	Yingchao	
郭金龙	Guo Jinlong	
金人庆	Jin Renqing	
阴囊	Scrotum	
阴蒂	Clitoris	
陈炳德	Chen Bingde	
达赖喇嘛	Dalai Lama	
氯胺酮	Ketamine	
学潮	Campus upheaval	
黄菊	Huang	
李鹏	Li Peng	
彭丽媛	Peng Liyuan	
温家宝	Wen Jiabao	
吴邦国	Wu Bangguo	
习近平	Xi Jinping	
曾庆红	Zeng Qinghong	
暴政	Tyranny	
达赖喇嘛	Dalai Lama	

傅政华	FuZhengHua	
傅政华	FuZhengHua	
共产党	Communist party	
共产党	Communist party	
共产党	Communist party	
热比娅·卡德尔	Rebiya Kadeer	
吴邦国	Wu Bangguo	
游行	Parade	
政变	Coup	
政变	Coup	
朱镕基	Zhu	
自焚	Self-burning	
江泽民	Jiang	
吕加平	Ping Lu Jiaguo	
暴乱	Riot	
罗干	Luo Gan	
阴唇	Labia	
支那	Shina	
六合彩	LOTTO	
六合彩	LOTTO	
淫妇	Adulteress	
阴户	Vagina	
阴精	Sex fluid	
阴门	Vulva	
淫浪	Yinlang	

Protected articles on Hudong Baike

Article title + Hudong link	Machine translation	Human translation / notes (if field is blank, translation to be confirmed)
苍井空	Sora Aoi	porn star
蔡英文	Tsai Ing-wen	Taiwanese politician
草榴社区	Grass garnet Community	Caoliu online forum
吴淑珍	Wu Shu-chen	wife of Chen Shui-bian
维基百科	Wikipedia	.
江泽民	Jiang Zemin	.
李长春[中央政治局常委]	Li Changchun [Politburo Standing Committee]	.
李克强	Li Keqiang	.
陈良宇	Chen Liangyu	Shangai pol dismissed for corruption
赌博	Gambling	.
朴善英	Park Sun-Young (Luna)	Korean pop star
习仲勋	Xi Zhongxun	father of Xi Jinping
习近平	Xi Jinping	.
令计划	Orders Scheme	
侯德建	Hou Jian	
俞正声	Yu	
刘云山[1]	Liu Yunshan [1]	
刘宾雁	Liu Binyan	
刘延东	Liu Yandong	

刘淇	Liu Qi	
刘晓波	Liu Xiaobo	
吴邦国	Wu Bangguo	
周永康	Zhou Yongkang	
回良玉	Hui Liangyu	
太子党	Princelings	
张德江	Zhang	
张高丽	Zhang Gaoli	
徐才厚	Xu Caihou	
文字狱	Inquisition	
文字狱	Inquisition	
方励之	Fang Lizhi	
方励之	Fang Lizhi	
无国界记者	Reporters Without Borders	
李源潮	Li Yuanchao	
李瑞环	Li Ruihuan	
李瑞环	Li Ruihuan	
李长春[中央政治局常委]	Li Changchun [Politburo Standing Committee]	
李鹏[国务院前总理]	Li [former Prime Minister of the State Council]	
柴玲	Chai Ling	
江泽民	Jiang	
汪洋[中央政治局委员]	Wang Yang [Politburo member]	
温家宝	Wen Jiabao	
王乐泉	Mr Wang	
王兆国	Wang	

王岐山	Wang Qishan	
王有才	Wang Youcai	
王乐泉	Mr Wang	
王立军[原重庆市副市长]	Wang Lijun [former vice mayor of Chongqing]	
王维林	Wang Weilin	
绝食	Fast	
罗干	Luo Gan	
罗干	Luo Gan	
习近平	Xi Jinping	
胡耀邦	Hu Yaobang	
胡锦涛	Hu	
莫言	Mo Yan	
华国锋	Hua Guofeng	
薄熙来	Bo Xilai	
诺贝尔和平奖	Nobel Peace Prize	
贾庆林	Jia Qinglin	
贺国强	He Guoqiang	
贾庆林	Jia Qinglin	
转法轮	Zhuan Falun	
达赖喇嘛	Dalai Lama	
郭伯雄	Guo	
陈光诚	Chen Guangcheng	
马英九	Ma Ying-jeou	
阿沛·阿旺晋美	Ngapoi Ngawang Jigme	
《中华人民共和国宪法》	Constitution of the People	

610办公室	610 Office	
李鹏[国务院前总理]	Li [former Prime Minister of the State Council]	
中国共产党	Communist Party of China	
世界维吾尔代表大会	World Uyghur Congress	
世界维吾尔代表大会	World Uyghur Congress	
彭丽媛	Peng Liyuan	
中华人民共和国	People	
中华民国	Republic of China	
中华民国	Republic of China	
人权	Human rights	
何清涟	He Qinglian	
共产党	Communist party	
华国锋	Hua Guofeng	
大跃进	Great Leap Forward	
李登辉[台湾地区前领导人]	Lee Teng-hui [former leader of Taiwan]	
毛泽东	Mao Zedong	
民主	Democracy	
汶川大地震	Wenchuan earthquake	
汶川大地震	Wenchuan earthquake	
草泥马	Mud horse	
百度	Baidu	
口交	Oral sex	
刘德华	Andy Lau	
金正恩[金正日之子]	Kim Jong-un [Kim Jong Il's son]	
微信	Micro-channel	

《一路向西》	All the way west.	
百度	Baidu	
淘宝网	Taobao	
性交	Sexual intercourse	
淘宝网	Taobao	
刘德华	Andy Lau	
林志玲	Ling	
文化大革命	Cultural Revolution	
母亲节	Mother's Day	
性交体位	Sexual intercourse position	
金瓶梅[古代小说]	Golden Lotus [Ancient Novel]	
周恩来	Zhou Enlai	
法轮	Falun	
林彪	Lin Biao	
肛交	Anal	
金正日	Kim Jong Il	
陈光标	Chen cursor	
刘志军	Liu Zhijun	
互动百科：联系我们	Interactive Encyclopedia: Contact Us	
射精	Ejaculation	
毛新宇	Mao Xinyu	
屌丝	Cock wire	
李雪主	Master Li Xue	
人体艺术	Body Art	
刘少奇	Liu Shaoqi	
胡春华	Hu Chunhua	

邓力群	Deng Liqun	
包皮	Foreskin	
朱镕基	Zhu	
广末凉子	Hirosue	
中华人民共和国	People	
妈祖	Matsu	
三级片	Three pieces	
维基百科	Wikipedia	
上户彩	Aya Ueto	
高虎城	Gao	
脱衣舞	Striptease	
朴槿惠	Park Geun-hye	
马馼	Ma Wen	
AV[成人影片]	AV [Adult Video]	
大阴唇	Labia majora	
崔智友	Choi Ji-woo	
北京地铁	Beijing Subway	
坂本真绫	Sakamoto Maaya	
大麻	Marijuana	
孙艺珍	Son Ye Jin	
推背图	Tui Bei Tu	
王立军[原重庆市副市长]	Wang Lijun [former vice mayor of Chongqing]	
陈希同	Chen Xitong	
三唑仑	Triazolam	
奥特曼	Altman	

邱毅	Chou Yi	
六七暴动	1967 riots	
女上位	Female host	
慈禧太后	Empress Dowager Cixi	
杨开慧	Kaihui	
王丹	Wang Dan	
高干子弟	Among the children	
李丽珍	Loretta (Loletta)	aka Racel Lee (actress)
中南海	Zhongnanhai	
中国共产党	Communist Party of China	
刘晓波	Liu Xiaobo	
推背图	Tui Bei Tu	
朱德	Zhu De	
翁虹	Yvonne Yung	
胡德平	Hu Deping	
邓朴方	Deng Pufang	
做爱	Sex	
储波	Chu Bo	
冷溶	Leng Rong	
列确	Legqog	
吗啡	Morphine	
何厚铧	Edmund Ho	
何鲁丽	He Luli	
俞正声	Yu Zhengsheng	
厉无畏	Li Wuwei	

天安门	Tiananmen Square	
姜异康	Jiang Yikang	
姜恩柱	Jiang	
孟学农	Meng Xuenong	
成思危	Cheng Siwei	
戴相龙	Dai Xianglong	
文昌星	Wenchang star	
曾培炎	Zeng Peiyan	
朱云来	Mr Zhu	
李肇星	Li Zhaoxing	
杨白冰	Yang Bai Bing	
梁保华	Liang Baohua	
中华人民共和国中央军事委员会	PRC Central Military Commission	
向巴平措	Qiangba	
法轮	Falun	
淫水	Sexual secretion	
由喜贵	Xigui	
胡星斗	Mr Hu	
藏独	Tibetan separatist	
郭金龙	Guo Jinlong	
金人庆	Jin Renqing	
陈炳德	Chen Bingde	
隗福临	Kui Fulin	
龙应台	Lung Ying-tai	
AV[成人影片]	AV [Adult Video]	

K粉	K powder	
黄菊	Huang	
GCD	GCD	
藏族青年大会	Tibetan Youth Congress	
达赖喇嘛	Dalai Lama	
刘永清	Liu Yongqing	
罗干	Luo Gan	

This last table displays a list of all the nine characters or less (or four double-byte Chinese characters) keywords which successfully matched to an entry on Wikipedia, but when searched on Hudong or Baidu Baike, returned no matching results. As mentioned previously, the matching algorithm is still extremely simple, so some of these keywords may indeed have appropriate entries on Hudong or Baidu Baike. For instance, the table reports that the movie 戀夏500日 (500 Days of Summer) is not found on Baidu or Hudong. However, some manual digging uncovers that the movie is listed under the title 《和莎莫的500天》 and the Wikipedia one must be a Taiwan or Hong Kong-specific one. Thus, the below will no doubt contain false positives for terms it failed to find due to flawed searching on our part. Thus, the below is merely an initial stab at trying to uncover the content gaps in the two *baike* and will be refined in the coming months.

The first column is the keywords used to test, the third is the title of the Wikipedia article that was found (which may sometimes vary from the initial keyword depending on how Wikipedia interprets the keyword and handles the redirecting to what it thinks is the proper article), and the last two columns indicate whether or not that Baidu or Hudong located a matching article for the keyword.

Not surprisingly, a number of the entries appear to be related to Taiwan, which makes sense since the largest share of Wikipedia China readers and contributors come from Taiwan while Hudong and Baidu are primarily mainland China users. Thus, a number of Taiwan-related articles in Wikipedia China may not to considered noteworthy to mainland users, explaining why they are not found in Hudong or Baidu. However, others are more clearly omitted for political or prurient reasons.

The same caveat about the machine translations from above applies here. **Do not disseminate the machine translations without first verifying if they are correct. We are in the process of confirming the translations in the coming weeks and months.**

Articles found on Wikipedia China but not on Hudong or Baidu Baike

Keyword searched + Wikipedia link	Machine translation of keyword	Wikipedia title found	Wikipedia title machine translation	Human translation + notes (if field is blank, translation to be confirmed)	On Baidu?	On Hudong?
草榴社区	Grass garnet Community	草榴社区	Grass garnet Community	Caoliu online community; forum involved with discussing Wenzhou train crash	no	
丁文雄	Dingwen Xiong	丁文雄	Dingwen Xiong	expelled Vietnamese politician?	no	no
六四事件	June 4 incident	六四事件	June 4 incident	Tiananmen Sq, 1989	no	no
满族	Manchu	满族	Manchu	ethnic minority	no	
法輪功	Falun Gong	法轮功	Falun Gong	religion	no	no
齡記書店	Ling Kee	齡記書店	Ling Kee	small HK bookstore	no	no
反共主义	Anti-communism	反共主义	Anti-communism	.	no	
研勤科技	Maction Technologies	研勤科技	Maction Technologies	Taiwan navigation company	no	no
PAPAGO!	PAPAGO!	研勤科技	Maction Technologies	.		no
蔡瑞月	Tsai Jui-yueh	蔡瑞月	Tsai Jui-yueh	Taiwanese dance pioneer	no	
莫莉花	Molly Flowers	莫莉花	Molly Flowers	Chinese dissident	no	no
曾偉恩	Zeng Wayne	曾偉恩	Zeng Wayne	Taiwanese baseball player	no	no
千甲車站	One thousand A station	千甲車站	One thousand A station		no	no

顺阳范氏	Shun Yang Fan	顺阳范氏	Shun Yang Fan			no
徐寧濉	Seo Young-hee	徐寧熙	Seo Young-hee	Korean actress	no	no
發正念	Fazhengnian	發正念	Fazhengnian	Falun Gong meditation	no	no
許瑜真	Xu Yu True	許瑜真	Xu Yu True		no	no
世博車站	Expo Station	千甲車站	One thousand A station		no	no
竹科車站	Hsinchu Science Park Station	新莊車站 (新竹市)	Shinjo Station (Hsinchu, Taiwan)		no	no
南阳范氏	Nanyang Fan	顺阳范氏	Shun Yang Fan		no	no
痤弦	Ya Xian	痤弦	Ya Xian		no	
北西摩島	North Seymour Island	北西摩島	North Seymour Island		no	no
平行計算	Parallel Computing	并行计算	Parallel Computing		no	no
维权运动	Rights movement	维权运动	Rights movement		no	no
鍾嘉欣	Linda	鍾嘉欣	Linda		no	no
維權運動	Rights movement	中国的人权维护运动	Maintenance of China’s human rights movement	Weiquan movement: Chinese lawyers and activists defending civil rights	no	no
崔真理	CUI Truth	崔真理	CUI Truth		no	
丁子霖	Ding Zilin	丁子霖	Ding Zilin	leader of Tiananmen Mothers	no	
七一遊行	July march	七一大遊行	The July 1 march	annual Hong Kong pro-democracy march	no	no
七不讲	Seven do not speak	七不讲	Seven do not speak		no	

严家其	Yan Jiaqi	严家其	Yan Jiaqi		no	no
令谷	Order Valley	令谷	Order Valley		no	no
侯赛因江	Hussein Jiang	侯赛因江	Hussein Jiang		no	no
八九学运	1989 student movement	六四事件	June incident		no	no
六四	Sixty-four	六四事件	June incident		no	no
刘宾雁	Liu Binyan	刘宾雁	Liu Binyan		no	
司徒华	SZETO	司徒華	SZETO		no	
司徒華	SZETO	司徒華	SZETO		no	
吴弘达	Harry Wu	吳弘達	Harry Wu		no	no
嚴家其	Yan Jiaqi	严家其	Yan Jiaqi		no	no
姜维平	Jiang Weiping	姜维平	Jiang Weiping		no	
孙文广	Sun Wenguang	孙文广	Sun Wenguang		no	
封从德	Feng Congde	封从德	Feng Congde	Chinese dissident	no	no
封從德	Letters from Germany	封从德	Letters from Germany		no	no
廖亦武	Liao Yiwu	廖亦武	Liao Yiwu		no	
张培莉	Zhangpei Li	张培莉	Zhangpei Li		no	
我的奋斗	Mein Kampf	我的奋斗	Mein Kampf			no
新疆独立	Xinjiang independence	新疆独立运动	Xinjiang independence movement		no	no
新疆獨立	Xinjiang independence	新疆独立运动	Xinjiang independence movement		no	no
李洪志	Li Hongzhi	李洪志	Li Hongzhi		no	no
柴玲	Chai Ling	柴玲	Chai Ling		no	
梁海怡	Liang Haiyi	梁海怡	Liang Haiyi		no	no

楊建利	Yang Jianli	楊建利	Yang Jianli		no	no
法轮功	Falun Gong	法轮功	Falun Gong		no	no
洛桑森格	Lobsang Senge	洛桑森格	Lobsang Senge		no	
洪哲勝	Hong Zhesheng	洪哲勝	Hong Zhesheng		no	no
温云松	Wenyun Song	温云松	Wenyun Song		no	
王军涛	Wang Juntao	王军涛	Wang Juntao		no	no
王力雄	Wang Lixiong	王力雄	Wang Lixiong		no	
王炳章	Wang Bingzhang	王炳章	Wang Bingzhang		no	
王维林	Wang Weilin	王维林	Wang Weilin		no	
盘古乐队	Pangu Band	盤古樂團	Pangu Orchestra		no	no
盛雪	Sheng Xue	盛雪	Sheng Xue		no	no
真善忍	Truthfulness	法轮功	Falun Gong		no	no
胡温体制	Hu-Wen regime	胡温体制	Hu-Wen regime		no	no
自由门	Freegate	自由门	Freegate		no	
袁红冰	Yuan Hongbing	袁红冰	Yuan Hongbing		no	no
辛灝年	Xin Hao years	辛灝年	Xin Hao years		no	no
金盾工程	Golden Shield Project	金盾工程	Golden Shield Project		no	
陈光诚	Chen Guangcheng	陈光诚	Chen Guangcheng		no	
韩东方	Han Dongfang	韩东方	Han Dongfang		no	no
高智晟	Gao Zhisheng	高智晟	Gao Zhisheng		no	
魏京生	Wei Jingsheng	魏京生	Wei Jingsheng		no	no
鮑彤	Bao Tong	鮑彤	Bao Tong		no	no
鮑彤	Bao Tong	鮑彤	Bao Tong		no	no
黎安友	Nathan	黎安友	Nathan		no	no
08宪章	Charter 08	零八宪章	Charter 08		no	no

花园软件	Garden Software	花园网	Garden Fence		no	no
西厢计划	West Wing plan	西厢计划	West Wing plan		no	no
GFW	GFW	防火长城	Great Firewall		no	no
唐柏桥	Tang Baiqiao	唐柏桥	Tang Baiqiao		no	no
幽灵网	Ghost Network	幽灵网	Ghost Network		no	
噶倫赤巴	Kalon Tripa	噶倫赤巴	Kalon Tripa		no	no
北京之春	Beijing Spring	北京之春	Beijing Spring		no	no
华夏文摘	China Digest	华夏文摘	China Digest		no	no
吾尔开希	Wuerkaixi	吾尔开希	Wuerkaixi		no	no
吾爾開希	Wuerkaixi	吾尔开希	Wuerkaixi		no	no
四五行動	AFA	四五行動	AFA		no	
敏感人士	Sensitive persons	持不同政见者	Political dissident		no	no
日本地理	Japan Geography	日本地理	Japan Geography			no
曹長青	Cao Changqing	曹長青	Cao Changqing		no	no
袁紅冰	Yuan Hongbing	袁红冰	Yuan Hongbing		no	no
零八宪章	Charter 08	零八宪章	Charter 08		no	no
民主女神	Goddess of Democracy	民主女神	Goddess of Democracy		no	no
牛博网	Bullog Network	牛博网	Bullog Network		no	
佔領中環	Occupy Central	佔領中環	Occupy Central			no
大紀元	The Epoch Times	大紀元	The Epoch Times		no	no
比特克	Beattock	比特克	Beattock		no	no
色情	Porn	色情	Porn		no	
花花公子	Playboy	花花公子	Playboy		no	
许志永	Xu Zhiyong	许志永	Xu Zhiyong		no	
谭作人	Tan Zuoren	谭作人	Tan Zuoren		no	

钓鱼岛	Diaoyu Islands	釣魚臺	Diaoyu Islands		no	
Anti-CNN	Anti-CNN	四月网	April networks		no	no
SARS事件	SARS incident	SARS事件	SARS incident		no	
鄧佳華	Deng Jiahua	鄧佳華	Deng Jiahua		no	
夜勤病棟	Night differential Ward	夜勤病棟	Night differential Ward		no	
鋼鐵人3	Iron Man 3	鋼鐵人3	Iron Man 3		no	no
性交体位	Sexual intercourse position	性交体位	Sexual intercourse position		no	
北川杏树	Kitagawa apricot	北川杏树	Kitagawa apricot		no	
肛交	Anal	肛交	Anal		no	
鋼鐵人2	Iron Man 2	鋼鐵人2	Iron Man 2			no
屍行者	Corpse Walker	陰屍路	Inferi Road		no	no
井川ゆい	Yui Igawa	井川由衣	Yui Igawa		no	no
戀曲寫真	Love Song Photo	戀曲寫真	Love Song Photo		no	no
S-Cute	S-Cute	S-Cute	S-Cute		no	no
北条麻妃	Hojo Ma Princess	北条麻妃	Hojo Ma Princess			no
山口昇	Yamaguchi Noboru	山口昇	Yamaguchi Noboru			no
朱音唯	Zhu Yin Wei	朱音唯	Zhu Yin Wei		no	
十四K	Fourteen K	14K	14K		no	
IU (歌手)	IU (singer)	IU (歌手)	IU (singer)		no	no
飢餓遊戲	The Hunger Games	飢餓遊戲	The Hunger Games			no
北川瞳	Hitomi Kitagawa	北川瞳	Hitomi Kitagawa			no
後藤理沙	Goto Risa	後藤理沙	Goto Risa		no	
沈夢辰	Shen Meng Chen	沈夢辰	Shen Meng Chen			no

愛情女僕	Love maid	愛情女僕	Love maid			no
爆漫王。	Bakuman.	爆漫王。	Bakuman.			no
金泰奭	Jintai pliable	金泰奭	Jintai pliable		no	no
成人电影	Adult Movies	成人电影	Adult Movies		no	
陰莖口交	Penis Fellatio	陰莖口交	Penis Fellatio		no	no
口內射精	Mouth ejaculation	口內射精	Mouth ejaculation			no
习明泽	Xi Mingze	习近平	Xi Jinping		no	no
郝明莉	Hao Mingli	郝明莉	Hao Mingli		no	
废赵贵人	Waste Zhaogui people	廢貴人趙氏	Waste elegant Zhao		no	no
zh-yue	zh-yue	粵语	Cantonese		no	no
成人影片	Adult Video	色情片	Porn		no	
緘默權	Right to silence	緘默權	Right to silence		no	
李鍾碩	Li Zhongshuo	李鍾碩	Li Zhongshuo		no	no
水菜麗	Mizuna Korea	水菜麗	Mizuna Korea			no
畢書盡	Bi books do	畢書盡	Bi books do		no	no
V City	V City	V City	V City		no	no
张宏堡	Zhang Hongbao	张宏堡	Zhang Hongbao		no	
成人网站	Adult Websites	成人网站	Adult Websites		no	
少年阿賓	Junior Abin	少年阿賓	Junior Abin		no	no
玩命關頭6	Fast and Furious 6	玩命關頭6	Fast and Furious 6		no	no
西門夜說	Simon said the night	西門夜說	Simon said the night		no	no
飛虎出征	Tiger expedition	飛虎出征	Tiger expedition			no
暗殺教室	Assassination classroom	暗殺教室	Assassination classroom			no

瘋狂假面	Crazy Mask	瘋狂假面	Crazy Mask		no	
顏射	Bukkake	颜射	Bukkake		no	
陳艾熙	Chen Yixi	陳艾熙	Chen Yixi		no	
神鎗狙擊	Gunslinger sniper	神鎗狙擊	Gunslinger sniper		no	no
蘇門達臘	Sumatra	蘇門答臘	Sumatra		no	
北欧神话	Norse mythology	北欧神话	Norse mythology		no	
Tsubomi	Tsubomi	蕾 (AV女優)	Lei (AV Actress)		no	no
黃山怡	Huangshan Yi	黃山怡	Huangshan Yi		no	
大橋未久	Not long Bridge	大橋未久	Not long Bridge		no	no
章粹吾	Zhang Cui I	章粹吾	Zhang Cui I		no	
中島春雄	Haruo Nakajima	中島春雄	Haruo Nakajima		no	no
呂宇俊	Lv Yujun	呂宇俊	Lv Yujun		no	no
大埔事件	Tai event	大埔事件	Tai event		no	no
嘉賢臺	Spectacle Taiwan	嘉賢臺	Spectacle Taiwan		no	no
電動扶梯	Escalators	電動扶梯	Escalators		no	
遺落戰境	Border war left behind	遺落戰境	Border war left behind		no	
8月	August	8月	August		no	
BTOB	BTOB	BTOB	BTOB		no	
Speak Now	Speak Now	爱的告白	Confessions of Love			no
Garie	Garie	Garie	Garie		no	no
傳說生物	Legend of biological	传说生物列表	Legend biological List		no	no
YouPorn	YouPorn	YouPorn	YouPorn		no	no
劉以豪	Liu Hao	劉以豪	Liu Hao			no
橘子熊	Orange Bear	橘子熊	Orange Bear		no	no

汪蘇瀧	Wang Su Long River	汪蘇瀧	Wang Su Long River		no	
希志愛野	Xizhi Aino	希志愛野	Xizhi Aino			no
冠军歌王	King of Champions	冠军歌王	King of Champions		no	
歐陽妮妮	Ouyang Nini	歐陽妮妮	Ouyang Nini		no	
飛虎II	Tiger II	飛虎II	Tiger II		no	
黃之鋒	Huang Feng	黃之鋒	Huang Feng		no	no
劉寅娜	Liu Yinna	劉寅娜	Liu Yinna		no	no
蘇怡賢	Su Yixian	蘇怡賢	Su Yixian		no	
蘇有朋	Alec	蘇有朋	Alec		no	
董智森	Dong Zhisen	董智森	Dong Zhisen		no	no
氣體行星	Gas planets	氣體巨行星	Gas giant planets		no	no
&	&	&	&			no
农历七月	Lunar July	农历七月	Lunar July		no	no
唐慧案	Hui Tang case	唐慧案	Hui Tang case			no
關澤新一	Sekisawa new one	關澤新一	Sekisawa new one		no	no
陳伊凌	Chen Yiling	陳伊凌	Chen Yiling		no	no
史托雅	Shi Tuoya	史托雅	Shi Tuoya		no	no
无码视频	Uncensored Video	无码视频	Uncensored Video		no	no
藤原瞳	Hitomi Fujiwara	藤原瞳	Hitomi Fujiwara		no	
馮·迪索	Von Paradiso	馮·迪索	Von Paradiso		no	no
上海帮	Shanghai to help	上海帮	Shanghai to help		no	no
宏盟集團	Omnicom Group	宏盟集團	Omnicom Group		no	
羅嘉仁	Luo Jiaren	羅嘉仁	Luo Jiaren		no	
T.O.P.	T.O.P.	T.O.P.	T.O.P.			no

戀夏500日	500 days of Summer	戀夏500日	500 days of Summer	Taiwan or HK title? Baidu and Hudong have it as 和莎莫的500天.	no	no
江國慶案	Chiang Kuo-ching case	江國慶案			no	no
犬種列表	Breed List	犬種列表	Breed List		no	no
蘇笠汶	Su Li Wen	蘇笠汶	Su Li Wen		no	
M Club	M Club	M Club			no	no
啟晴邨	Kai Ching Estate	啟晴邨	Kai Ching Estate		no	
大浦安娜	Oura Anna	大浦安娜				no
畢·彼特	Bi Pitt	畢·彼特			no	
真愛配方	Love Recipe	真愛配方				no
豪斯的头	House’s head	豪斯的头			no	
新楓之谷	New MapleStory	新楓之谷			no	no
石首事件	Stone’s first event	石首事件	Stone’s first event		no	
萌學園五	Meng Gakuen five	萌學園五異界對決			no	no
AV (電影)	AV (movie)	AV (電影)			no	no
TARI TARI	TARI TARI	TARI TARI	TARI TARI			no
五毛党	Fifty Cent Party	网络评论员	Network commentator		no	no
佛地魔	Voldemort	伏地魔			no	
俞强声	YU strong voice	俞强声	YU strong voice		no	no
徐朱玄	Xu Zhu Xuan	徐朱玄	Xu Zhu Xuan		no	no
李艺真	Li Yi really	李艺真	Li Yi really			no
海女姑娘	Female sea girl	海女小天	Female sea small day		no	

結界女王	Enchantment Queen	結界女王	Enchantment Queen			no
金汎	Gold Pan	金汎	Gold Pan			no
AKBINGO!	AKBINGO!	AKBINGO!			no	no
june 4th	june 4th	六四事件	June incident		no	no
五毛蛋	Fifty Cent egg	五毛蛋争议	Fifty Cent egg controversy		no	no
师 (军队)	Division (military)	师 (军队)	Division (military)		no	no
色情演員	Porn actor	色情演員			no	no
陰道口	Vagina	陰道口	Vagina		no	
黃紫盈	Huang Ziying	黃紫盈	Huang Ziying			no
LEE HI	LEE HI	李遐怡	Li Yi-ya		no	no
The xx	The xx	The xx				no
一世代	A generation	一世代	A generation		no	no
山達基	Scientology	山達基	Scientology		no	
幸運⭐星	Lucky ? Star	幸運⭐星				no
朴珊德拉	??? pull	朴珊德拉			no	no
歐陽娜娜	Ouyang Nana	歐陽娜娜	Ouyang Nana		no	
水嶋杏美	Mizushima apricot America	水嶋杏美	Mizushima apricot America		no	no
特攻聯盟	Kick Ass	特攻聯盟	Kick Ass			no
蘇強文	Su Qiang Wen	蘇強文	Su Qiang Wen		no	
金宇彬	Jin Yubin	金宇彬	Jin Yubin			no
金明洙	Jin Mingzhu	金明洙				no
韩寒困境	Han dilemma	韩寒被质疑造假事件	Han was challenged fraud case		no	no
刺蝟男孩	Hedgehog Boy	刺蝟男孩			no	no

吳青峯	Wu's	吳青峯	Wu's			no
城巴85P線	Citybus 85P line	城巴85P線	Citybus 85P line		no	
張友驊	ZHANG Hua	張友驊			no	
心有花	Heart with flowers	心有花	Heart with flowers			no
溫子仁	Wenzi Ren	溫子仁	Wenzi Ren		no	no
虎II坦克	Tiger II tanks	虎II坦克			no	
護聖院宮	Hushengyuangong	護聖院宮	Hushengyuangong		no	
A BEST	A BEST	A BEST				no
BTOOOM!	BTOOOM!	BTOOOM!	BTOOOM!		no	no
FOX娛樂台	FOX Showbiz	FOX娛樂台	FOX Showbiz		no	
Futtminx	Futtminx	Tuttminx	Tuttminx		no	no
于正昇	In n l	于正昇				no
南波杏	Namba apricot	南波杏				no
咲-Saki-	Saki-Saki-	咲-Saki-	Saki-Saki-			no
成人動畫	Adult animation	成人動畫	Adult animation		no	
新發邨	San Fat Estate	新發邨	San Fat Estate		no	no
朴修映	Pu Xiu Ying	朴修映	Pu Xiu Ying		no	no
李遐怡	Li Yi-ya	李遐怡	Li Yi-ya		no	no
村上里沙	Risa Murakami	竹內紗里奈				no
林百貨	Forest department	林百貨			no	
真白希實	Really white Heshbon	真白希實	Really white Heshbon		no	
绘色千佳	Chika color painting	绘色千佳	Chika color painting			no
聯發科技	MediaTek	聯發科技	MediaTek		no	no
西鐵綫	West Rail Line	西鐵綫	West Rail Line			no

鍾麗緹	Christy	鍾麗緹	Christy		no	no
Waking Up	Waking Up	甦醒 (共和世代專輯)				no
台灣論壇	Taiwan Forum	台灣論壇	Taiwan Forum		no	
天兵公園	Creation of the park	天兵公園	Creation of the park		no	no
崔東昱	Cuidong Yu	崔東昱	Cuidong Yu		no	no
手交	Handjob	手交	Handjob			no
燒夷彈	Napalm bombs	凝固汽油弹	Napalm		no	
瑪雅曆	Mayan calendar	瑪雅曆	Mayan calendar			no
直通运行	Run through	直通运行	Run through		no	
網路性交	Internet Sex	網路性交	Internet Sex		no	no
若宮莉那	That Wakamiya Li	若宮莉那				no
蔡雪瑩	Cai Xueying	蔡雪瑩	Cai Xueying		no	no
超級偶像8	Super Idol 8	超級偶像8	Super Idol 8		no	no
趙權	Jo Kwon	趙權	Jo Kwon		no	
長沙灣站	Cheung Sha Wan Station	長沙灣站				no
高志超	Gao Chao	高志超	Gao Chao		no	
九评	Nine Commentaries	九評	Nine Commentaries		no	no
共军	Troops	共軍	Troops		no	no
共狗	Of the dog	共匪	Bandits		no	no
兽交	Bestiality	動物戀	Animal Love		no	no
刘荻	Liu Di	刘荻	Liu Di		no	
办证	Accreditation	办假证	Do false		no	no
吞精	Swallow	口內射精	Mouth ejaculation		no	

咸片	Salty piece	色情片	Porn		no	no
妓男	Male prostitutes	男妓	Prostitutes		no	
射颜	Shot Yan	颜射	Bukkake		no	no
巨奶	Giant milk	巨乳	Boobs		no	no
人獸交	Bestiality	動物戀	Animal Love		no	
刘贤斌	Liu Xianbin	刘贤斌	Liu Xianbin		no	no
办假证	Do false	办假证	Do false		no	no
北高联	North Gaolian	北京高校学生自治联合会	Beijing College Student Government Association		no	no
华雪和	China snow and	华雪和	China snow and		no	no
吴扬明	Wu Yang Ming	吴扬明	Wu Yang Ming		no	
呼喊派	Shouters	呼喊派	Shouters		no	
圣辉洛	Sheng Hui Luo	聖輝洛	Sheng Hui Luo		no	no
姬胜德	Ji Shengde	姬胜德	Ji Shengde		no	
小鸡巴	Small Cock	阴茎	Penis		no	no
干你娘	Dry your mother	幹你娘	Dry your mother		no	no
庞维勒	Pangwei Le	西奥多·庞维勒	Theodore Pangwei Le			no
徐勤先	Xu Qin first	徐勤先	Xu Qin first			no
怯口活	Cowardly mouth live	倒口	Inverted mouth		no	no
李常受	Lee	李常受	Lee		no	
李志绥	Li Zhisui	李志绥	Li Zhisui		no	
4月03日	03-Apr	4月03日	03-Apr		no	no
71遊行	71 march	七一大遊行	The July 1 march		no	no
89运动	89 Sports	六四事件	June incident		no	no

一中一台	One China, one Taiwan	一邊一國	Each side		no	no
一党专制	One-party dictatorship	一党制	One-party system		no	no
三人性交	Threesome	三人性行为	Trio sex		no	no
五四事件	May incident	五四运动	Fourth Movement		no	no
人造阴道	Artificial vagina	人工陰道	Artificial vagina		no	no
八四事件	Eighty-four event	香港的士罷駛事件	Hong Kong taxi strikes		no	no
北京政权	Beijing regime	中华人民共和国	People		no	no
卢多维西	Ludovisi	卢多维西	Ludovisi		no	no
地方召会	Local churches	召會	Church		no	no
大陸當局	Chinese authorities	中華人民共和國政府	PRC Government		no	no
天下围攻	World siege	百萬人民倒扁運動	Million people fell flat movement		no	no
小江调水	Small river diversion	小江调水	Small river diversion		no	no
巴爾幹炮	Balkan gun	M61火神式機砲	M61 Vulcan cannon		no	no
希望之声	Sound of Hope	希望之声国际广播电台	Sound of Hope Radio		no	no
干你老母	Dry your mother	幹你娘	Dry your mother		no	no
新闻管制	Press Control	新闻自由	Freedom of the press		no	no
毛匪	Hair bandit	毛泽东	Mao Zedong		no	no
油外泄	Oil leakage	油外洩	Oil leakage		no	no
淫水	Sexual secretion	陰道分泌液	Vaginal fluids		no	

獸交	Bestiality	動物戀	Animal Love		no	no
王文怡	Wang Wenyi	王文怡	Wang Wenyi		no	no
維基揭密	Wiki Revealed	維基解密	WikiLeaks		no	no
维基揭密	Wiki Revealed	維基解密	WikiLeaks		no	no
网络封锁	Internet blockade	互联网审查	Internet censorship		no	no
美分党	Cents Party	网络特工	Network agents		no	
肏你妈	Fuck your mother	幹你娘	Dry your mother		no	no
肏你媽	Fuck your mother	幹你娘	Dry your mother		no	no
胡海峰	Hu Haifeng	胡海峰	Hu Haifeng		no	
胡石根	Hu Shi root	胡石根	Hu Shi root		no	no
茉莉革命	Jasmine Revolution	茉莉花革命	Jasmine Revolution		no	no
草你妈	Grass your mother	幹你娘	Dry your mother		no	no
蒙汗药	Drugged	蒙汗药	Drugged		no	
虐肛	Anal abuse	肛交	Anal		no	no
血房地图	Blood House Map	血房地图	Blood House Map		no	no
被立王	Be king	被立王	Be king		no	no
裙底春	Skirts spring	裙底風光	Skirts scenery		no	no
裙底風光	Skirts scenery	裙底風光	Skirts scenery		no	no
裙底风光	Skirts scenery	裙底風光	Skirts scenery		no	no
铁血旗	Jagged flag	十八星旗	Eighteen Star flag		no	no
颜射	Bukkake	颜射	Bukkake		no	
鸡巴	Cock	陽具	Penis		no	no
大纪元	The Epoch Times	大紀元	The Epoch Times		no	no
共惨党	Total miserable	共匪	Bandits		no	no

	party					
温影帝	Wen winner	温家宝	Wen Jiabao		no	no
June 4th	June 4th	六四事件	June incident		no	no
藏青会	TYC	西藏青年大會	Tibetan Youth Congress		no	
程映虹	Cheng Yinghong	程映虹	Cheng Yinghong		no	no
动态网	Dynamic Network	自由门	Freegate			no
胡海清	Hu Haiqing	胡海清	Hu Haiqing		no	no
九評	Nine Commentaries	九評	Nine Commentaries		no	no
民阵	SNF	民間人權陣線	Civil Human Rights Front		no	no
土共	Total soil	香港親共人士	Hong Kong's pro-people			no
团派	Tuanpai	團派	Tuanpai		no	no
希望之聲	Sound of Hope	希望之声国际广播电台	Sound of Hope Radio		no	no
辛子陵	Xin Ziling	辛子陵	Xin Ziling		no	
学运	Student movement	學生運動	Student movement		no	no
赵连海	Zhao drafted	赵连海	Zhao drafted		no	
北京当局	Beijing authorities	北京政府	Beijing Government		no	no
高自联	High self-Union	北京高校学生自治联合会	Beijing College Student Government Association		no	no
何德普	He Depu	何德普	He Depu		no	no
红色恐怖	Red Terror	红色恐怖	Red Terror		no	no
李旺阳	Li Wangyang	李旺陽事件	Li Wangyang event		no	no

民主墙	Democracy Wall	民主墙	Democracy Wall		no	no
大陆当局	Chinese authorities	中華人民共和國政府	PRC Government		no	no
反攻大陆	Retake the mainland	反攻大陸	Retake the mainland		no	no
少年阿宾	Junior Abin	少年阿賓	Junior Abin		no	no
国殇之柱	Pillar of Shame	國殤之柱	Pillar of Shame		no	
八九學運	1989 student movement	六四事件	June incident		no	no

NEXT STEPS

As the matching mechanism appears to work, expanding the keyword list is necessary to start building a larger database for analysis. However, at some point, as the keywords become ever more obscure and less common, we'll no doubt see diminishing returns after expanding the keyword list too far.

Using the other variables we've collected will also make for interesting analysis later on. For example, models could be created to predict what kinds of articles are most likely to be protected, or for spotting articles that have abnormally small numbers of edits and article lengths. A discussion of what users are trying to express when they vote a particular entry as helpful would also be worthwhile. For instance, the most liked article on Hudong was one for a banking executive while the top ones on Baidu are mostly pop-related: Jacky Wu, Michael Jackson, the Japanese manga Detective Conan, singer Yang Mi, Jay Chou, and… Mao Zedong, with over 236,000 thumbs ups–though all of them were blown out of the water by the Korean group 2NE1, whose Baidu Baike article was liked nearly 2 million times. What are users trying to say when they mark something as helpful by clicking the thumbs up icon on each *baike* service? For instance, what can we conclude from the fact that more people have classified Bo Xilai's entry on Baidu and Hudong as "useful" than they have for Xi Jinping's? Should we think of "likes" for particular pages as an alternative way netizens can actively or passively signal their endorsement of a topic's existence or article's sentiment even if the content is drastically different from a more "neutral" source like Wikipedia?

By looking not only at what content and data doesn't exist on *baike,* but also at the content that does, this project will investigate what knowledge and information is fit for public display. If articles are shorter on Hudong and Baidu, what information do they carry? Does this information reveal anything about the authors' intentions? By examining the topics and articles that are left visible in these *baike* and considering the motivations behind those who seek out, view, edit, and approve of these articles, this project hopes to offer a more nuanced view of the typical narratives about censorship in China. Trying to understand what sorts of expressions netizens are making via these online encyclopedias, despite whatever censorship might be taking place, is as interesting as the potential censorship itself. This project will hopefully push us to once again consider the many complexities when discussing information control in environments where oversight of content has been decentralized to companies and users–an environment which makes it increasingly harder to identify traditional instances of censorship.

1. If you search for most general topics on Baidu, links to Baidu images, search, videos, shopping, bulletin

boards, answers, and its own encyclopedia are among the top results. For example, here's a search for 手机 (cell phone): Baidu links make up more than half of the first page of results, including a hit to Baike. When you search for celebrities and political figures, Baidu Baike is again always among the top results, if not the very first one. Hudong is often not found on the first page of results–if at all. Here's a search for Xi Jinping. Notice the message at the top in bold (根据相关法律法规和政策，部分搜索结果未予显示。) stating that the results have been filtered. Those that are returned are all state media links (Xinhua, People's Daily, China Daily, CNTV/CCTV)–except for Baike which inherits the top spot. Even for searches of people that are not filtered, Baike still almost always gets top billing and Hudong is often relegated to later in the results–though often not due to any failure on Hudong's part. For example, the same search for 周恩来 (Zhou Enlai) on Google has both Hudong and Baidu Baike in the top three results (Wikipedia garners the top spot). All this is to say that while giving preference to one's own services is not some unprecedented online tactic (see the outcry in 2010 when Google tweaked its searches to prioritize YouTube and other internal Google-owned content above other comparable websites), it does seem like there is something inherently baked into Baidu's search algorithm that either intentionally or unintentionally penalizes Hudong's articles compared to Baidu Baike and others. However, more rigorous examination is needed before a conclusion about how real or serious that penalty can be made.

2. For example, here's an article on cloud seeding that was featured on Hudong on August 7, 2013. It's section on the costs of cloud seeding are copied word for word in Baidu Baike on the same topic (starts with "一种是用飞机把干冰等冷却剂撒播到云中"). If you look at the version histories, the Hudong version that contains this section predates Baidu by nearly four years. For Wikipedia copying, take a look at each respective encyclopedia's entry on 利玛窦 (Matteo Ricci). Here's a comparison of what is shared between Wikipedia and Hudong today, and Wikipedia and Baidu Baike today. As you can see from the matching bits, Wikipedia and Hudong share much greater overlap, and if you go through the historical records, you can identify that Hudong added the majority of such material in the May 11, 2008 edit. If you take the Apr 27, 2008 version of the article from Wikipedia and compare the two, you'll notice just how closely the two mirror each other. The same can be done with Baidu Baike: a comparison of the differences between the Aug 29, 2006 Wikipedia article on Matteo Ricci and the Sept 6, 2006 Baidu Baike one. The overlap is stark and damning.

Again, this is not rigorous at all, merely a few cases that I have stumbled upon and require more rigorous examination before deciding whether the claims of plagiarism by Baidu and Hudong are unfounded, incidental (a few novice users unclear about citations), or systemic (either intentionally directed or through willful ignorance and refusal to correct it by those in charge).

About the Author

Jason Q. Ng is the Google Policy Fellow at The Citizen Lab and author of *Blocked on Weibo*, a book about censorship and sensitive topics in Chinese social media. You can read more of his work at his blog or follow him on Twitter.