

The English Wikipedia Database

Paolo Tamagnini
Benedetta Checcarelli



WIKIPEDIA
The Free Encyclopedia

page
<ul style="list-style-type: none"> page_id INT page_namespace INT page_title VARCHAR(255) page_restrictions TINYBLOB page_is_redirect TINYINT page_is_new TINYINT page_random DOUBLE page_touched BINARY(14) page_links_updated VARBINARY(14) page_latest INT page_len INT page_content_model VARBINARY(32) page_lang VARBINARY(35)
Indexes

pagelinks
<ul style="list-style-type: none"> pl_from INT pl_from_namespace INT pl_namespace INT pl_title VARCHAR(255)
Indexes

category
<ul style="list-style-type: none"> cat_id INT cat_title VARCHAR(255) cat_pages INT cat_subcats INT cat_files INT
Indexes

imagelinks
<ul style="list-style-type: none"> il_from INT il_from_namespace INT il_to VARCHAR(255)
Indexes

categorylinks
<ul style="list-style-type: none"> cl_from INT cl_to VARCHAR(255) cl_sortkey VARBINARY(230) cl_sortkey_prefix VARCHAR(255) cl_timestamp TIMESTAMP cl_collation VARBINARY(32) cl_type ENUM(...)
Indexes

image
<ul style="list-style-type: none"> img_name VARCHAR(255) img_size INT img_width INT img_height INT img_metadata MEDIUMBLOB img_bits INT img_media_type ENUM(...) img_major_mime ENUM(...) img_minor_mime VARBINARY(100) img_description VARCHAR(767) img_user INT img_user_text VARCHAR(255) img_timestamp VARBINARY(14) img_sha1 VARBINARY(32)
Indexes

```
mysql> create table page_sample as  
-> SELECT *  
-> FROM page  
-> where rand() <= 0.01  
-> limit 30000;
```

CREATING A NEW TABLE
WITH A SAMPLE OF 30K
RANDOM TUPLES FROM THE
PAGE TABLE

```
mysql> select *  
-> from page_sample  
-> INTO OUTFILE 'page_30k.csv'  
-> FIELDS TERMINATED BY ','  
-> ENCLOSED BY ''''  
-> LINES TERMINATED BY '\n';
```

EXPORTING THE SAME
TABLE IN A CSV FILE

USING THE SAME TABLE TO EXPORT ALL THE
CATEGORIES OF THE PAGES IN THE SAMPLE

```
mysql> select c.*  
-> from category c  
-> where c.cat_title in (  
->         select cl.cl_to  
->         from page_sample ps, categorylinks cl  
->         where cl.cl_from = ps.page_id )  
-> INTO OUTFILE 'category_30k.csv'  
-> FIELDS TERMINATED BY ','  
-> ENCLOSED BY '"'  
-> LINES TERMINATED BY '\n';
```

EXPORTING RELATIVE PAGELINKS, IMAGELINKS,
CATEGORYLINKS AND IMAGES IN THE SAME WAY

```
mysql> select pl.*  
-> from pagelinks pl, page_sample ps  
-> where pl.pl_from = ps.page_id  
-> INTO OUTFILE 'pagelinks_30k.csv'  
-> FIELDS TERMINATED BY ','  
-> ENCLOSED BY '"'  
-> LINES TERMINATED BY '\n';
```

page
page_id INT
page_namespace INT
page_title VARCHAR(255)
page_restrictions TINYBLOB
page_is_redirect TINYINT
page_is_new TINYINT
page_random DOUBLE
page_touched BINARY(14)
page_links_updated VARBINARY(14)
page_latest INT
page_len INT
page_content_model VARBINARY(32)
page_lang VARBINARY(35)
Indexes

categorylinks
cl_from INT
cl_to VARCHAR(255)
cl_sortkey VARBINARY(230)
cl_sortkey_prefix VARCHAR(255)
cl_timestamp TIMESTAMP
cl_collation VARBINARY(32)
cl_type ENUM(...)
Indexes

category
cat_id INT
cat_title VARCHAR(255)
cat_pages INT
cat_subcats INT
cat_files INT
Indexes

image
img_name VARCHAR(255)
img_size INT
img_width INT
img_height INT
img_metadata MEDIUMBLOB
img_bits INT
img_media_type ENUM(...)
img_major_mime ENUM(...)
img_minor_mime VARBINARY(100)
img_description VARCHAR(767)
img_user INT
img_user_text VARCHAR(255)
img_timestamp VARBINARY(14)
img_sha1 VARBINARY(32)
Indexes

imagelinks
il_from INT
il_from_namespace INT
il_to VARCHAR(255)
Indexes

pagelinks
pl_from INT
pl_from_namespace INT
pl_namespace INT
pl_title VARCHAR(255)
Indexes







Exporting
SQL queries

MySQL

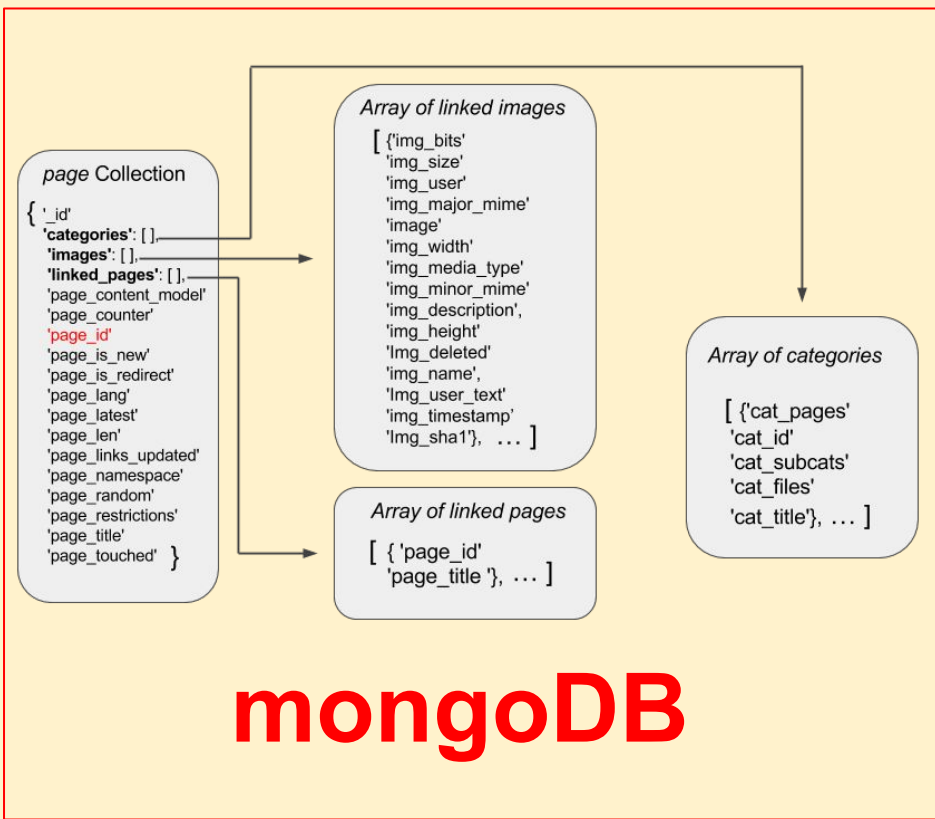
category_30k.csv
categorylinks_30k.csv
image_30k.csv
imagelinks_30k.csv
page_30k.csv
pagelinks_30k.csv

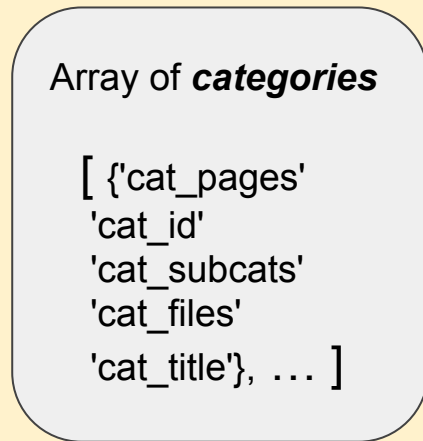
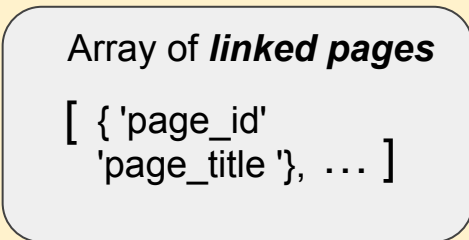
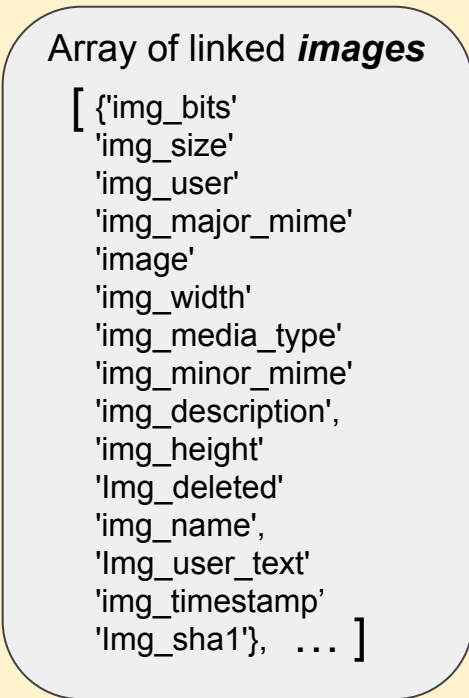
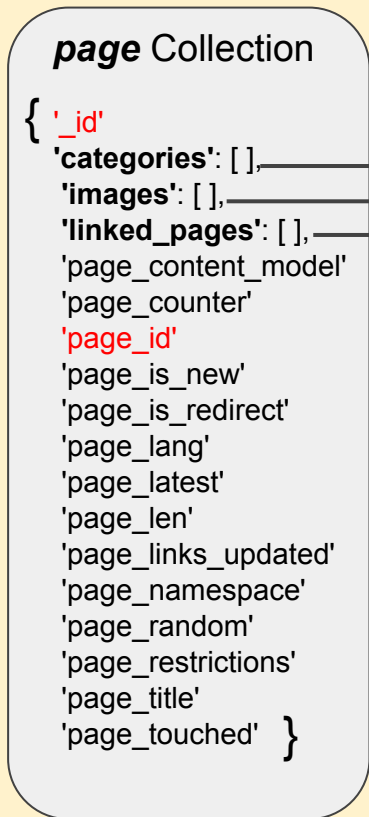
Option 1:

just one collection with
embedded arrays of documents

 category_30k.csv
 categorylinks_30k.csv
 image_30k.csv
 imagelinks_30k.csv
 page_30k.csv
 pagelinks_30k.csv

pymongo






```
> db.page.createIndex( { 'page_id': 1 } )
{
  "createdCollectionAutomatically" : false,
  "numIndexesBefore" : 1,
  "numIndexesAfter" : 2,
  "ok" : 1
}
```

```
> db.page.find(
... {"$and":[
...   { "categories.cat_title": {"$regex" : ".*_War_.*"}},
...   {"page_title": {"$regex" : ".*_submarine_.*"} }
... ] },
... {"page_title":1,"_id":0} )
```

```
{ "page_title" : "Soviet_submarine_K-8" }
{ "page_title" : "German_submarine_U-110_(1940)" }
{ "page_title" : "German_submarine_U-573" }
```

```
> db.page.distinct("page_namespace")
```

```
[ 0.0,  
  1.0,  
  4.0,  
100.0,  
  2.0,  
  3.0,  
  5.0,  
  6.0,  
  8.0,  
10.0,  
11.0,  
14.0,  
15.0,  
  7.0,  
  9.0,  
101.0,  
13.0,  
12.0 ]
```

Wikipedia data structure			
Namespaces			
Subject namespaces		Talk namespaces	
0	(Main/Article)	Talk	1
2	User	User talk	3
4	Wikipedia	Wikipedia talk	5
6	File	File talk	7
8	MediaWiki	MediaWiki talk	9
10	Template	Template talk	11
12	Help	Help talk	13
14	Category	Category talk	15
100	Portal	Portal talk	101
108	Book	Book talk	109
118	Draft	Draft talk	119
446	Education Program	Education Program talk	447
710	TimedText	TimedText talk	711
828	Module	Module talk	829
2300	Gadget	Gadget talk	2301
2302	Gadget definition	Gadget definition talk	2303

```
> db.page.aggregate([
...   {"$match": {"page_touched": {"$gt": 20170000000000 } }},
...   {"$group": {"_id": "$page_namespace",
...               "count": {"$sum": 1}}},
...   {"$sort" : { "count" : -1 } }])
```

```
{ "_id" : 0, "count" : 15309 }
{ "_id" : 1, "count" : 2781 }
{ "_id" : 14, "count" : 723 }
{ "_id" : 6, "count" : 501 }
{ "_id" : 3, "count" : 327 }
{ "_id" : 2, "count" : 125 }
{ "_id" : 4, "count" : 117 }
{ "_id" : 10, "count" : 116 }
{ "_id" : 15, "count" : 38 }
{ "_id" : 5, "count" : 23 }
{ "_id" : 11, "count" : 17 }
{ "_id" : 100, "count" : 13 }
{ "_id" : 101, "count" : 4 }
{ "_id" : 12, "count" : 2 }
{ "_id" : 7, "count" : 1 }
{ "_id" : 9, "count" : 1 }
{ "_id" : 8, "count" : 1 }
{ "_id" : 13, "count" : 1 }
```

Wikipedia data structure			
Namespaces			
Subject namespaces		Talk namespaces	
0	(Main/Article)	Talk	1
2	User	User talk	3
4	Wikipedia	Wikipedia talk	5
6	File	File talk	7
8	MediaWiki	MediaWiki talk	9
10	Template	Template talk	11
12	Help	Help talk	13
14	Category	Category talk	15
100	Portal	Portal talk	101
108	Book	Book talk	109
118	Draft	Draft talk	119
446	Education Program	Education Program talk	447
710	TimedText	TimedText talk	711
828	Module	Module talk	829
2300	Gadget	Gadget talk	2301
2302	Gadget definition	Gadget definition talk	2303

```
> db.page.aggregate([
...   {"$unwind": "$images"},
...   {"$group":   { "_id": "$page_title",
...                 "images": {"$push": "$images.img_name"},
...                 "#_of_images": {"$sum": 1} }   },
...   {"$sort": {"#_of_images": -1}}])
```

```
{ "_id" : "Starfury", "images" : [ "B5FuryBadger.jpg", "B5FuryFork.jpg", "B5FuryHeavy.jpg",
"B5FuryLight.jpg", "B5FuryStealth.jpg", "B5FuryThunder.jpg" ], "#_of_images" : 6 }
{ "_id" : "And_All_That_Could_Have_Been", "images" : [ "Aatchb_dlx.jpg", "Aatchb_dvd.jpg",
"Aatchb_live.jpg", "Aatchb_still.jpg" ], "#_of_images" : 4 }
{ "_id" : "Anthem_(The_Wildhearts_song)", "images" : [ "Anthemwildheartscd1.jpg",
"Anthemwildheartscd2.jpg", "Anthemwildheartsvinyl.jpg" ], "#_of_images" : 3 }
{ "_id" : "Arkham_Asyllum", "images" : [ "Arkham-Asylum.jpg",
"Arkham_Asyllum_(Batman_-258_(October_1974)).jpg", "BATMAN_SHADOW_OF_THE_BAT_82.jpg" ], "#_of_images" : 3 }
{ "_id" : "Jacksonville_Jaguars", "images" : [ "2009-2012AFCS-UniformCombinations-JAX.PNG",
"AFCS-1995-1996-Uniform-JAX.PNG", "AFCS-Uniform-Combination-JAX.PNG" ], "#_of_images" : 3 }
{ "_id" : "Australian_two-dollar_coin", "images" : [ "Australian_$2_Coin.png",
"Australian_$2_Coin_2012_Remembrance.jpg" ], "#_of_images" : 2 }
[ ... ]
```

```

> db.page.aggregate([
...     {"$unwind": "$images"},
...     {"$group": {"_id": "$images.img_user_text",
...                 "unique_images":
...                     {"$addToSet": "$images.img_name"}}},
...     {"$project": {"_id": 1,
...                     'unique_images': 1,
...                     "#_of_images": { "$size": "$unique_images" } }},
...     {"$sort": {"#_of_images": -1}}])

{ "_id" : "Theo's Little Bot", "unique_images" : [ "1947_NZ_Test_team.jpg", ... ], "#_of_images" : 9 }
{ "_id" : "718 Bot", "unique_images" : [ "Alberta_Fed_logo.png", ... ], "#_of_images" : 8 }
{ "_id" : "Minsk59", "unique_images" : [ "B5FuryStealth.jpg", ... ], "#_of_images" : 6 }
{ "_id" : "J Greb", "unique_images" : [ "ACswim.jpg", ... ], "#_of_images" : 4 }
{ "_id" : "Spineback", "unique_images" : [ "Anthemwildheartsvinyl.jpg", ... ], "#_of_images" : 3 }
{ "_id" : "Pais", "unique_images" : [ "AppleShare_IP_Migration_screenshot.png", ... ], "#_of_images" : 3 }
{ "_id" : "Fuzzy510", "unique_images" : [ "AFRICA_CONFIDENTIAL_LOGO.jpg", ... ], "#_of_images" : 3 }
{ "_id" : "NeoBatfreak", "unique_images" : [ "Aquamantitle.png", ... ], "#_of_images" : 3 }
[ ... ]

```

Option 2:

A collection for every table of
the old MySQL database

category_30k.csv
categorylinks_30k.csv
image_30k.csv
imagelinks_30k.csv
page_30k.csv
pagelinks_30k.csv

pymongo

page {}	category {}	imagelinks {}
page_id INT	cat_id INT	il_from INT
page_namespace INT	cat_title VARCHAR(255)	il_from_namespace INT
page_title VARCHAR(255)	cat_pages INT	il_to VARCHAR(255)
page_restrictions TINYBLOB	cat_subcats INT	
page_is_redirect TINYINT	cat_files INT	Indexes
page_is_new TINYINT	Indexes	
page_random DOUBLE		pagelinks {}
page_touched BINARY(14)		pl_from INT
page_links_updated VARBINARY(14)		pl_from_namespace INT
page_latest INT		pl_namespace INT
page_len INT		pl_title VARCHAR(255)
page_content_model VARBINARY(32)		Indexes
page_lang VARBINARY(35)		
Indexes		
	image {}	
	img_name VARCHAR(255)	
	img_size INT	
	img_width INT	
	img_height INT	
	img_metadata MEDIUMBLOB	
	img_bits INT	
	img_media_type ENUM(...)	
	img_major_mime ENUM(...)	
	img_minor_mime VARBINARY(100)	
	img_description VARCHAR(767)	
	img_user INT	
	img_user_text VARCHAR(255)	
	img_timestamp VARBINARY(14)	
	img_sha1 VARBINARY(32)	
	Indexes	
categorylinks {}		
cl_from INT		
cl_to VARCHAR(255)		
cl_sortkey VARBINARY(230)		
cl_sortkey_prefix VARCHAR(255)		
cl_timestamp TIMESTAMP		
cl_collation VARBINARY(32)		
cl_type ENUM(...)		
Indexes		

mongoDB

```
db.page.createIndex( { 'page_id': 1 } )  
db.pagelinks.createIndex( { 'pl_from': 1, 'pl_title': 1 } )  
db.image.createIndex( { 'img_name': 1 } )  
db.imagelinks.createIndex( { 'il_from': 1, 'il_to': 1 } )  
db.category.createIndex( { 'cat_id': 1 } )  
db.categorylinks.createIndex( { 'cl_from': 1, 'cl_to': 1 } )
```



```
> var page_Star = db.page.find({"page_title":"Starfury"})
> var pageIds = page_Star.map(function(c) {return c.page_id;})
> db.imagelinks.find({il_from: {$in:pageIds}}, {il_to: 1,_id:0})
```

```
{ "il_to" : "B5FuryBadger.jpg" }
{ "il_to" : "B5FuryFork.jpg" }
{ "il_to" : "B5FuryHeavy.jpg" }
{ "il_to" : "B5FuryLight.jpg" }
{ "il_to" : "B5FuryStealth.jpg" }
{ "il_to" : "B5FuryThunder.jpg" }
```

```
> var user_Theo = db.image.find({"img_user_text":"Theo's Little Bot"})
> var images_of_Theo = user_Theo.map(function(c) {return c.img_name;})
> var page_Theo = db.imagelinks.find({il_to: {$in:images_of_Theo}})
> var pageIds_Theo = page_Theo.map(function(c) {return c.il_from;})
> db.page.find({page_id: {$in:pageIds_Theo}}, {page_title: 1,_id:0})
```

```
{ "page_title" : "Hercule Poirot" }
{ "page_title" : "Jacksonville Jaguars" }
{ "page_title" : "Kristallnacht" }
{ "page_title" : "Rockefeller_Center" }
{ "page_title" : "Arkham_Asyllum" }
{ "page_title" : "Let's_Make_a_Deal" }
{ "page_title" : "Abia_State" }
{ "page_title" : "Assault_on_Precinct_13_(1976_film)" }
{ "page_title" : "Don_Taylor_(cricketer)" }
```

```
mysql> select pl_title
```

```
-> from pagelinks
```

```
-> where pl_from in (
```

```
-> select page_id
```

```
-> from page
```

```
-> where page_title in (
```

```
-> select pl_title
```

```
-> from pagelinks
```

```
-> where pl_from in (
```

```
-> select page_id
```

```
-> from page
```

```
-> where page_title = "List_of_Mars-crossing_minor_planets"))));
```

```
+-----+
| pl_title |
+-----+
| (163364)_2002_OD20 |
| (277475)_2005_WK4 |
| (285263)_1998_QE2 |
| (29074)_5160_T-3 |
| (29076)_1972_TR8 |
| (35395)_1997_XM10 |
| (35397)_1997_YJ |
| (410777)_2009_FD |
| (52759)_1998_MW13 |
| (52761)_1998_MN14 |
| (53318)_1999_JV7 |
| (53320)_1999_JW8 |
| (6177)_1986_CE2 |
| (68949)_2002_QN6 |
| (68951)_2002_QH27 |
| (7888)_1993_UC |
| (89958)_2002_LY45 |
| (89960)_2002_ND35 |
| 101955_Bennu |
| 10301_Kataoka |
| 10303_Fr |®ret |
| 1035_Amata |
| 1037_Davidweilla |
| 1220_Crocus |
| 1222_Tina |
| 131_Vala |
| 133_Cyrene |
+-----+
```

```
27 rows in set (5 min 7.61 sec)
```

Telecommunications in Italy

From Wikipedia, the free encyclopedia

See also: *List of Italian telephone companies*, *List of radio*

Telephones - main lines in use:^[1] 20.031 million (2008)

Telephones - mobile cellular:^[1] 88.58 million (2008)

Telephone system:^[1] modern, well-developed, fast; fully automatic; ^[1] high-capacity cable and microwave radio relay trunk lines; ^[1] *domestic:* high-capacity cable and microwave radio relay trunk lines; ^[1] *international:* satellite earth stations - 3 Intelsat (with a total of 30 circuits); 21 submarine cables.

Radio broadcast stations:^[1] AM about 100, FM about 4,600

Radios: 50.5 million (1997)


Television broadcast stations:^[1] 358 (plus 4,728 repeaters)

Televisions: 30.5 million (1997)

Internet Hosts:^[1] 22.152 million (2009)

Internet users:^[1] 24.992 million (2008)

Country code (Top-level domain): **.it**

V · T · E	Country code top-level domains	[hide]
	ISO 3166-1	[hide]
<div><div><div><div>A .ac .ad .ae .af .ag .ai .al .am .ao .aq .ar .as .at .au .aw .ax .az</div><div>B .ba .bb .bd .be .bf .bg .bh .bi .bj .bm .bn .bo .br .bs .bt .bw .by</div><div>.bz C .ca .cc .cd .cf .cg .ch .ci .ck .cl .cm .cn .co .cr .cu .cv .cw .cx</div><div>.cy .cz D .de .dj .dk .dm .do .dz E .ec .ee .eg .er .es .et .eu F .fi</div><div>.fj .fk .fm .fo .fr G .ga .gd .ge .gf .gg .gh .gi .gl .gm .gn .gp .gq .gr</div><div>.gs .gt .gu .gw .gy H .hk .hm .hn .hr .ht .hu I .id .ie .il .im .in .io .iq</div><div>.ir .is .it J .je .jm .jo .jp K .ke .kg .kh .ki .km .kn .kp .kr .kw .ky .kz</div><div>L .la .lb .lc .li .lk .lr .ls .lt .lu .lv .ly M .ma .mc .md .me .mg .mh .mk</div><div>.ml .mm .mn .mo .mp .mq .mr .ms .mt .mu .mv .mw .mx .my .mz</div><div>N .na .nc .ne .nf .ng .ni .nl .no .np .nr .nu .nz O .om P .pa .pe .pf</div><div>.pg .ph .pk .pl .pm .pn .pr .ps .pt .pw .py Q .qa R .re .ro .rs .ru</div><div>.rw S .sa .sb .sc .sd .se .sg .sh .si .sk .sl .sm .sn .so .sr .ss .st .su</div><div>.sv .sx .sy .sz T .tc .td .tf .tg .th .tj .tk .tl .tm .tn .to .tr .tt .tv .tw .tz</div><div>U .ua .ug .uk .us .uy .uz V .va .vc .ve .vg .vi .vn .vu W .wf .ws</div><div>Y .ye .yt Z .za .zm .zw</div></div></div></div>		
Introduced	1987	
TLD type	Country code top-level domain	
Status	Active	
Registry	IT-NIC	
Sponsor	IIT-CNR	
Intended use	Entities connected with  Italy	
Actual use	Very popular in Italy Commonly used as a domain hack in English-speaking countries	
Registration restrictions	Must be a resident of an EU country to register. Domain name must be at least three characters long.	
Structure	Registration is permitted at second level; there are some third-level names beneath second-level labels, but these are not much used	
Documents	How to register	
Dispute policies	Dispute procedure	
Website	Italy NIC	

```
> var infos_TLC =  
...     db.page.find({"page_title":"Telecommunications_in_Italy"}).map(function(c){  
...                                     return c.page_id; });  
> var links_1 = db.pagelinks.find({pl_from: {$in:infos_TLC}}).map(function(c) {  
...                                     return c.pl_title; });  
> var ids_links_1 = db.page.find({page_title: {$in:links_1}}).map(function(c) {  
...                                     return c.page_id; });  
> db.pagelinks.find({pl_from: {$in:ids_links_1}},{pl_title:1,_id:0});
```

```
{ "pl_title" : ".ac" }  
{ "pl_title" : ".ad" }  
{ "pl_title" : ".ae" }  
{ "pl_title" : ".af" }  
{ "pl_title" : ".ag" }  
{ "pl_title" : ".ai" }  
{ "pl_title" : ".al" }  
{ "pl_title" : ".am" }  
{ "pl_title" : ".an" }  
[ ... ]
```