BNSI URLs Retrieval Software

ESSNET Big Data WP2

Rome, November 8

2 "P. Volov" Str., 1038 Sofia, Bulgaria, tel. +359
2 9857 729

Initial conditions

- The task
  - Finding websites of enterprises with 10 and more employees
- Software environment
  - Windows OS
  - MySQL, MSSQL, …
  - PHP, VB, Java, …
- Expertise of participants
  - MySQL, PHP

Initial data from Business Register

CSV file with 26836 businesses

- 20649 e-mails
- 2006 urls
- Addresses
- Phone numbers
- NACE codes
- …

MySQL Database

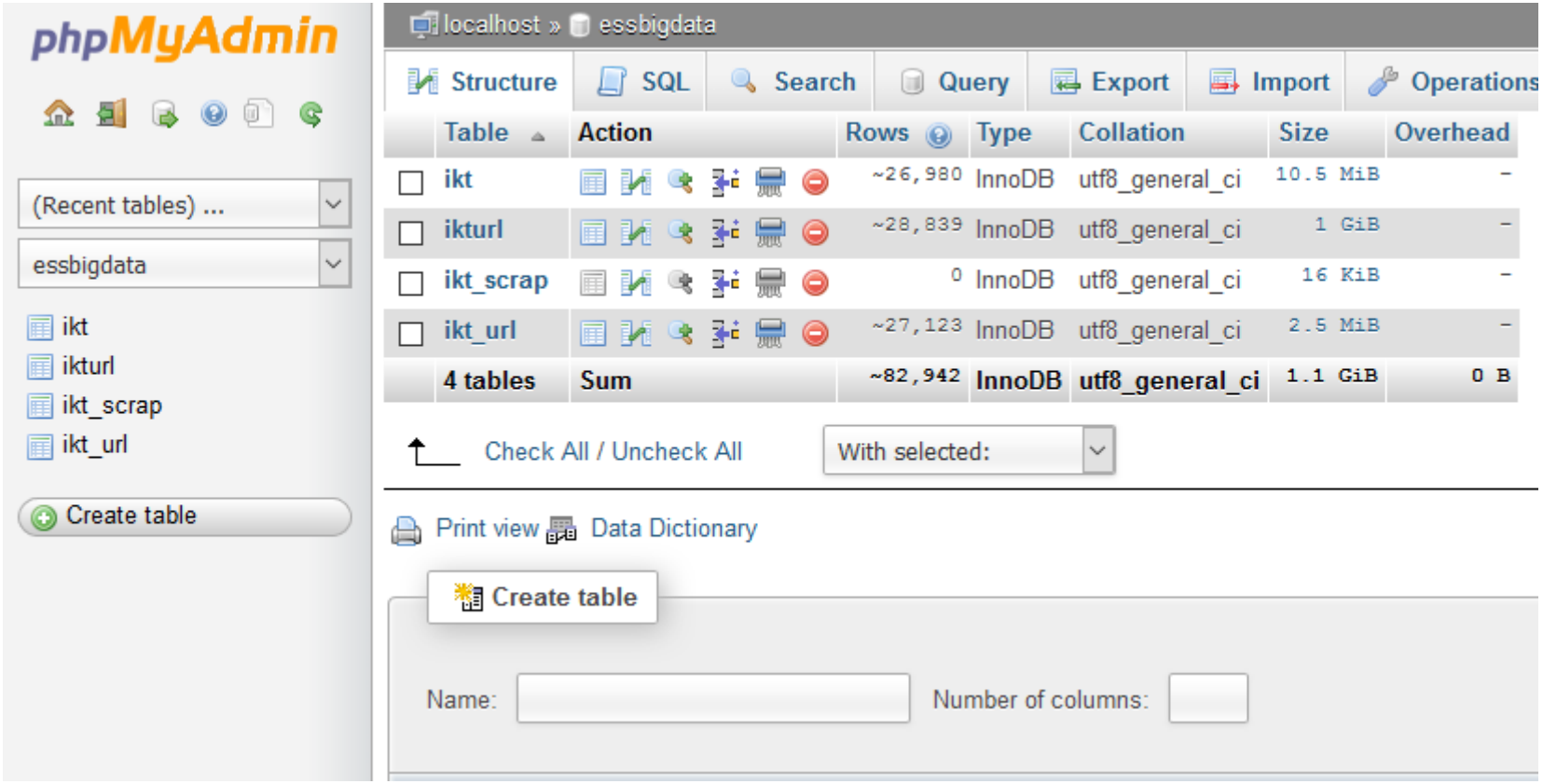
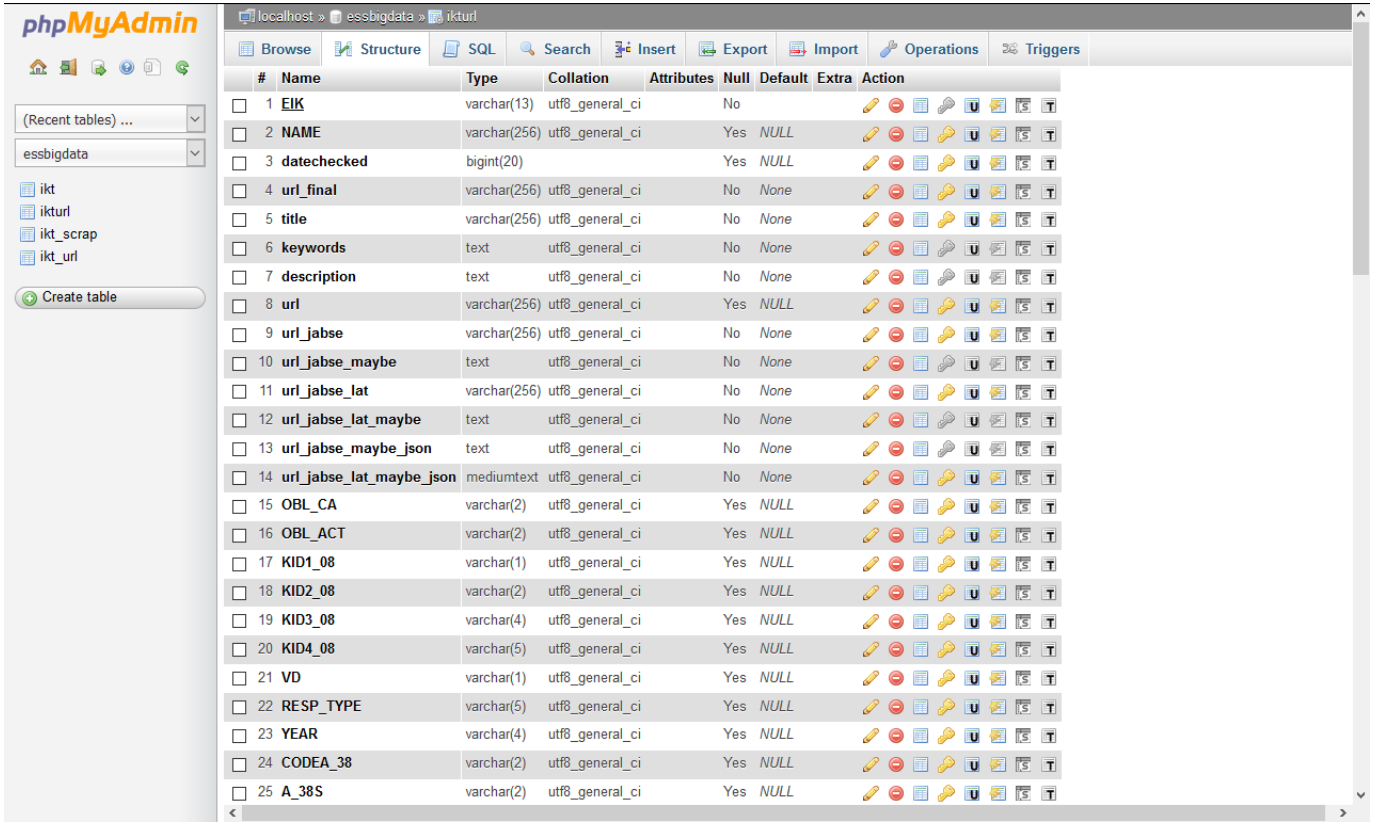
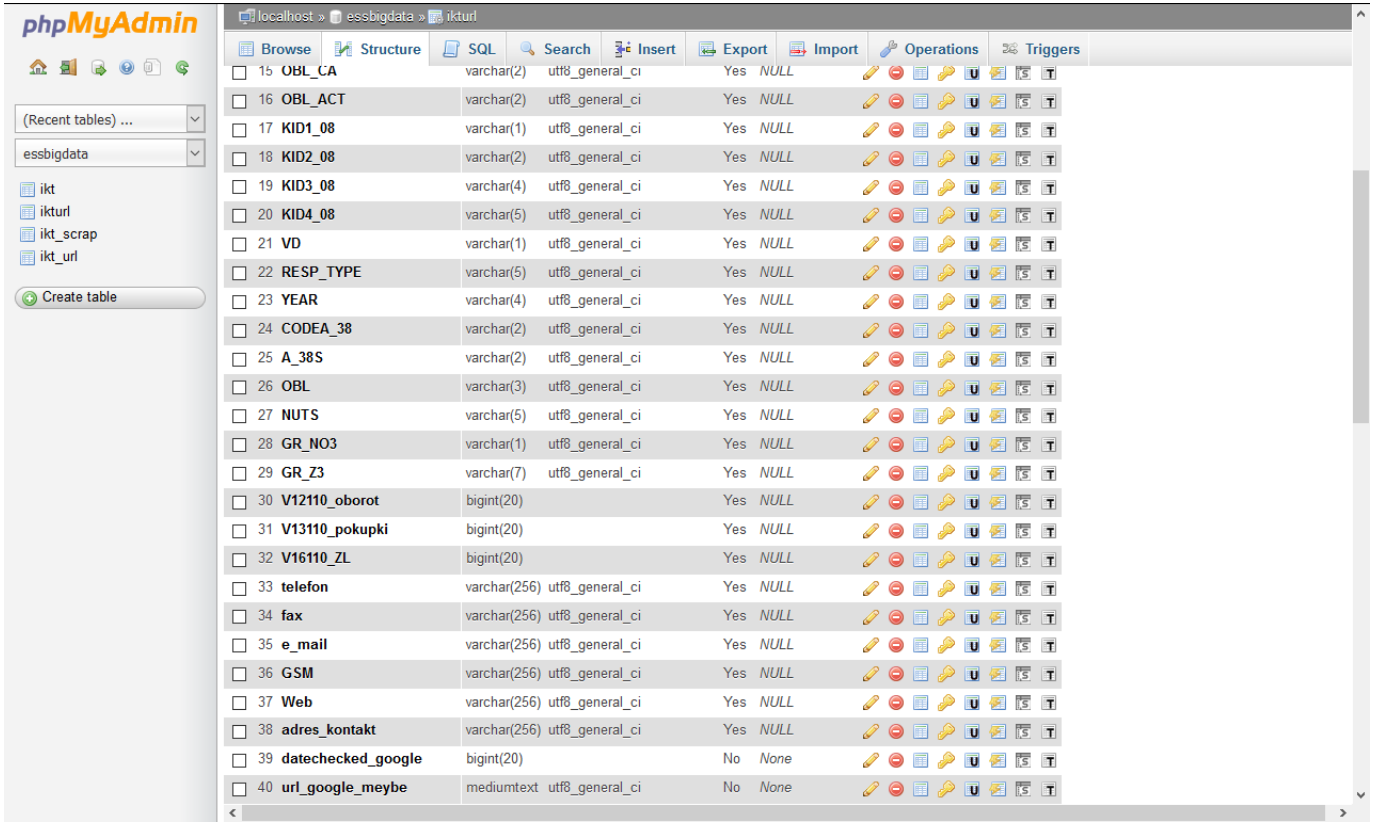


Table ikturl Structure (1)



#	Name	Type	Collation	Attributes	Null	Default	Extra	Action
1	EIK	varchar(13)	utf8_general_ci		No			
2	NAME	varchar(256)	utf8_general_ci		Yes	NULL		
3	datechecked	bigint(20)			Yes	NULL		
4	url_final	varchar(256)	utf8_general_ci		No	None		
5	title	varchar(256)	utf8_general_ci		No	None		
6	keywords	text	utf8_general_ci		No	None		
7	description	text	utf8_general_ci		No	None		
8	url	varchar(256)	utf8_general_ci		Yes	NULL		
9	url_jabse	varchar(256)	utf8_general_ci		No	None		
10	url_jabse_maybe	text	utf8_general_ci		No	None		
11	url_jabse_lat	varchar(256)	utf8_general_ci		No	None		
12	url_jabse_lat_maybe	text	utf8_general_ci		No	None		
13	url_jabse_maybe_json	text	utf8_general_ci		No	None		
14	url_jabse_lat_maybe_json	mediumtext	utf8_general_ci		No	None		
15	OBL_CA	varchar(2)	utf8_general_ci		Yes	NULL		
16	OBL_ACT	varchar(2)	utf8_general_ci		Yes	NULL		
17	KID1_08	varchar(1)	utf8_general_ci		Yes	NULL		
18	KID2_08	varchar(2)	utf8_general_ci		Yes	NULL		
19	KID3_08	varchar(4)	utf8_general_ci		Yes	NULL		
20	KID4_08	varchar(5)	utf8_general_ci		Yes	NULL		
21	VD	varchar(1)	utf8_general_ci		Yes	NULL		
22	RESP_TYPE	varchar(5)	utf8_general_ci		Yes	NULL		
23	YEAR	varchar(4)	utf8_general_ci		Yes	NULL		
24	CODEA_38	varchar(2)	utf8_general_ci		Yes	NULL		
25	A_38S	varchar(2)	utf8_general_ci		Yes	NULL		

Table ikturl Structure (2)



15	OBL_CA	varchar(2)	utf8_general_ci	Yes	NULL	
16	OBL_ACT	varchar(2)	utf8_general_ci	Yes	NULL	
17	KID1_08	varchar(1)	utf8_general_ci	Yes	NULL	
18	KID2_08	varchar(2)	utf8_general_ci	Yes	NULL	
19	KID3_08	varchar(4)	utf8_general_ci	Yes	NULL	
20	KID4_08	varchar(5)	utf8_general_ci	Yes	NULL	
21	VD	varchar(1)	utf8_general_ci	Yes	NULL	
22	RESP_TYPE	varchar(5)	utf8_general_ci	Yes	NULL	
23	YEAR	varchar(4)	utf8_general_ci	Yes	NULL	
24	CODEA_38	varchar(2)	utf8_general_ci	Yes	NULL	
25	A_38S	varchar(2)	utf8_general_ci	Yes	NULL	
26	OBL	varchar(3)	utf8_general_ci	Yes	NULL	
27	NUTS	varchar(5)	utf8_general_ci	Yes	NULL	
28	GR_NO3	varchar(1)	utf8_general_ci	Yes	NULL	
29	GR_Z3	varchar(7)	utf8_general_ci	Yes	NULL	
30	V12110_oborot	bigint(20)		Yes	NULL	
31	V13110_pokupki	bigint(20)		Yes	NULL	
32	V16110_ZL	bigint(20)		Yes	NULL	
33	telefon	varchar(256)	utf8_general_ci	Yes	NULL	
34	fax	varchar(256)	utf8_general_ci	Yes	NULL	
35	e_mail	varchar(256)	utf8_general_ci	Yes	NULL	
36	GSM	varchar(256)	utf8_general_ci	Yes	NULL	
37	Web	varchar(256)	utf8_general_ci	Yes	NULL	
38	adres_kontakt	varchar(256)	utf8_general_ci	Yes	NULL	
39	datechecked_google	bigint(20)		No	None	
40	url_google_meybe	mediumtext	utf8_general_ci	No	None	

Scripts

Index of /essBigData

Name	Last modified	Size	Description
Parent Directory		-	
geturl.php	25-Apr-2016 12:43	3.7K	
google_search.php	16-Jun-2016 15:05	1.9K	
info.php	30-Aug-2016 11:22	15K	
jabse_interface.php	26-Apr-2016 16:04	4.9K	
jabse_search.php	10-Jun-2016 13:02	4.6K	
list.php	17-Jun-2016 15:00	7.2K	

Common features of the search scripts

- Run in browser
- <meta http-equiv="Refresh" content="30">
- Timestamps

Script geturl.php

- Checks if the initial 2006 urls are real websites
- Constructs domain names from the 20649 e-mails
- Checks if the constructed domains are real websites
- Saves the results in the database
- Result - 7038 possible urls

Script jabse_search.php

- Uses automated search interface of http://www.jabse.com (jabse_interface.php)
- Get up to 10 search results from names in Bulgarian and the same for transliterated names in English
- Excluding from the search results the complex urls
- Saves the results in the database
- Result - 15638 results in Bulgarian and 16201 results in Latin

Jabse search interface

- 200 searches per hour
- Only in Bulgarian
- Jabse gives better results in English

Script google_search.php

- Uses Google search interface
- Get up to 10 search results from names in Bulgarian
- Saves the results in the database
- Result - 26829 sets of up to 10 search results

Google search interface

- 200 searches per day free
- 1000 searches for 5 EUR max 10000 per day
- 300 EUR for free searches on credit card registration

Script list.php

Database crawling interface, which displays the enterprises with characteristics and the urls search results and allows the user to choose the correct url of each enterprise

list.php.htm

Manual work done

- 26836 records were checked in 45 working days
- 600 records per work day
- 9809 urls were found
- 36.6 % of enterprises have websites

Results and statistics

info.php.htm

2 "P. Volov" Str., 1038 Sofia, Bulgaria, tel. +359
2 9857 729

The scripts

- Made to do the work in Bulgarian reality
- Not intended for different database table structure
- Hard coded labels in Bulgarian

Questions

- Are the scripts interesting for other countries?
- Should the scripts be configurable?
- Should the scripts be translatable?
- Is there a need for more search interfaces?

Thank you!

2 "P. Volov" Str., 1038 Sofia, Bulgaria, tel. +359
2 9857 729