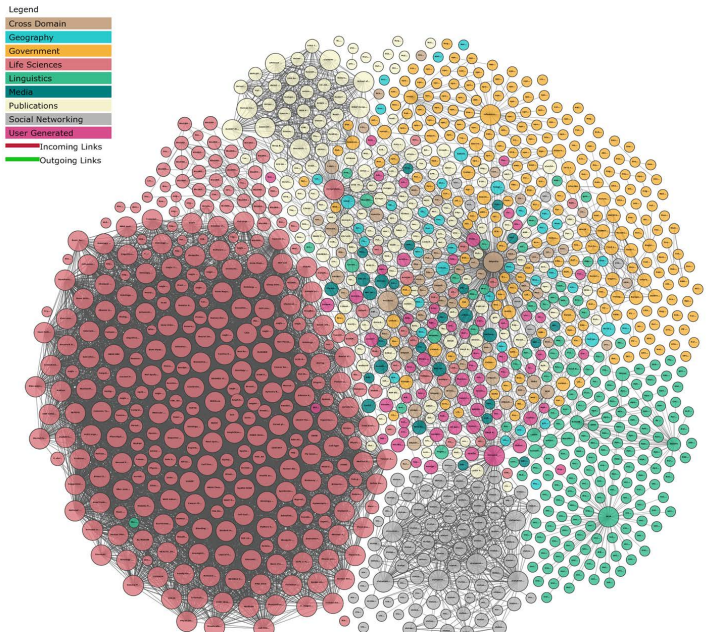
Characterizing user groups of DBpedia-NL through user log analysis

Frank Walraven - 2542405

What is DBpedia?

- Linked Data:
  - "*A method of publishing structured data so that it can be interlinked and become more useful through semantic queries*" (Ismayilov & Kontokostas, 2017)
- Freely available
  - SPARQL-endpoints
  - Local dumps of DBpedia
- Data from Wikipedia editions in 111 languages
- Users according to Official DBpedia Wiki:
  - Large companies
  - Libraries
  - Researchers
  - Web developers.
- Average of 7,343,939 hits per day (October 2016 - December 2017)



Dbpedia-NL

- DBpedia is divided into local chapters per language
  - Coordinated by DBpedia Internationalization Committee
- Focus on DBpedia-NL

Problem

- Improving DBpedia-NL
  - Who are the users and what do they expect?
- Limited data available on users

Research questions

- Research questions:
  - What is a good way of characterizing user groups that use DBpedia-NL?
  - What is DBPedia-NL currently used for?

Method - Summary

- Data-driven method
- Validated through survey
- Data-driven method:
  - User log analysis in combination with local DBpedia-NL dump to retrieve the popular categories of DBpedia-NL
  - Subjects and hierarchical relations to characterize the log entries into grouped categories
- Validation through survey
  - Data-driven method results ↔ Survey results
- International version of DBpedia

Method - Pre-processing user log

- Pre:
  - xxx.xx.xx.xxx - - [07/Dec/2017:03:34:23 +0100] "GET /resource/Roger_Cicero HTTP/1.1" 303 -

  IP — Date & time — Resource
- Post:
  - xx.xxx.xxx.xxx 07/Dec/2017:03:34:23 +0100 /resource/Roger_Cicero
- Office Access 2016 issues
  - Apostrophes
  - Brackets
  - Minus
  - “GET”
  - HTTP

Method - IP-address analysis

- Classify most used IP-addresses
  - General description, not conclusive

IP-address classes			
Class	**Start address**	**Finish address**	**Description**
A	0.0.0.0	126.255.255.255	Very large networks such as multinational companies
B	128.0.0.0	191.255.255.255	Large networks such as a college and ISP
C	192.0.0.0	223.255.255.255	Small to mid-sized companies
D	224.0.0.0	239.255.255.255	Multicast services
E	240.0.0.0	255.255.255.255	Reserved for experimental use

Method - IP-address analysis

- IP-address lookup
  - Useful information:
    - Host
    - Country
    - IP owner info
    - Domain owner info
  - IP: 68.180.xxx.xx
    Host: b1100 18.yse.yahoo.net
    Country: United States
  - Not always possible to determine if bot or not

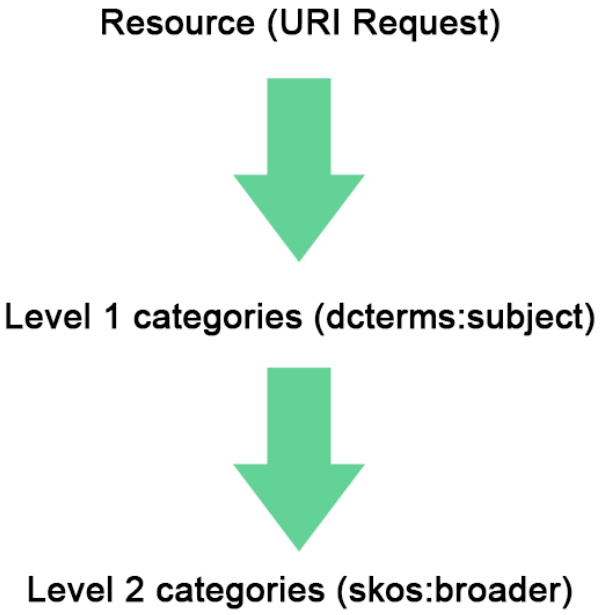
Results - IP-address analysis

- Top 30 most used IP-addresses:
  - 80% of IP-addresses Class A & bot
  - 16.67% unsure
  - 3.33% no bot
- Excluding Class A from most used IP-addresses:
  - 60% bots
  - 10% no bot
  - 30% unsure

#	IP-address	Frequency	IP-class	Bot?
1	68.180.xxx.xx	1981562	A	Yes (Yahoo)
2	17.142.xxx.xxx	782012	A	Yes (Apple)
3	130.251.xx.xxx	130780	B	Unsure
4	68.180.xxx.xx	123090	A	Yes (Yahoo)
5	78.46.xxx.xxx	82204	A	Unsure
6	46.229.xxx.xx	53190	A	Yes
7	46.229.xxx.xx	51658	A	Yes
8	46.229.xxx.xx	51314	A	Yes
9	46.229.xxx.xx	51286	A	Yes
10	46.229.xxx.xx	51192	A	Yes
11	46.229.xxx.xx	51164	A	Yes
12	46.229.xxx.xx	50690	A	Yes
13	46.229.xxx.xx	50656	A	Yes
14	46.229.xxx.xx	50414	A	Yes
15	46.229.xxx.xx	50326	A	Yes
16	46.229.xxx.xx	50146	A	Yes
17	46.229.xxx.xx	44992	A	Yes
18	46.229.xxx.xx	43850	A	Yes
19	138.201.xxx.xx	41336	B	Unsure
20	46.229.xxx.xx	41327	A	Yes
21	46.229.xxx.xx	40930	A	Yes
22	202.180.xx.xxx	32570	C	Unsure
23	46.229.xxx.xx	26408	A	Yes
24	68.180.xxx.xxx	25803	A	Yes (Yahoo)
25	216.244.xx.xxx	20600	C	Yes (Wowrack.com)
26	194.116.xx.xxx	19904	C	Unsure
27	17.142.xxx.xxx	18258	A	Yes (Apple)
28	216.244.xx.xxx	13996	C	Yes (Wowrack.com)
29	216.244.xx.xxx	13054	C	Yes (Wowrack.com)
30	160.45.xxx.xxx	12520	B	No (Stanford researchers)

Method - Categories

- Wikipedia Categories are represented in DBpedia using DCMI terms and the SKOS vocabulary.
  - “Dcterms:subject”
    - “Android_TV” resource:
      - “category-en:Google”
      - “category-en:Android_(operating_system)_software”
    - Level 1 category
  - “Skos:broader”
    - Connect subcategories and supercategories
    - Level 2 category



Method - Unused data

- Not all resources are linked to Level 1 or Level 2 categories
- Unused data Level 1 categories:
  - 787,039 unique URI requests in user log
  - 33.79% (265,926 URI requests) unused
- Unused data Level 2 categories
  - 70,892 unique Level 1 categories
  - 6.72% (4,766 Level 1 categories) unused

Method - Level 1 categories

- Extract Level 1 categories
  - List of URI Requests and their corresponding Level 1 categories
    - `SELECT level_1_categories.categories, uri_list.uri FROM level_1_categories INNER JOIN uri_list ON level_1_categories.uri = uri_list.uri;`
    - Sorted using Office Excel 2016
  - Normalized frequencies
    - High in-degree of resources

$$NormalizedFrequency = \frac{\#\ of\ UNIQUE\ resources\ in\ user\ log\ per\ Level\ 1\ Category}{Total\ \#\ of\ linked\ resource\ in\ DBpedia-NL\ per\ Level\ 1\ Category}$$

Result - Level 1 categories

- IUCN-status_niet_bedreigd
  - Rank 2
  - Normalized frequency: 27%
- Dier_uit_het_Palearctisch_gebied
  - Rank 10
  - Normalized frequency: 14%

Top 10 Level 1 categories (unique resources per Level 1 category)			
Level 1 category		Frequency	Normalized frequency
1	Amerikaanse_film	7440	74%
2	IUCN-status_niet_bedreigd	4565	27%
3	Dramafilm	3717	67%
4	Amerikaans_acteur	3158	83%
5	Nederlands_voetballer	2857	61%
6	Amerikaans_filmacteur	2819	81%
7	Amerikaans_televisieacteur	2500	80%
8	Komische_film	2125	69%
9	Pseudoniem	1928	73%
10	Dier_uit_het_Palearctisch_gebied	1913	14%

Method - Level 2 categories

- Extract Level 2 categories from Level 1 categories
  - SELECT Count(T.Field1) AS CountOfField1, T.Field2 FROM (SELECT DISTINCT user_log.ip, level_1_category_list.level1cat FROM level_1_category_list INNER JOIN user_log ON level_1_category_list.uri = user_log.uri GROUP BY level_1_category_list.level1cat, user_log.ip) AS T GROUP BY T.Field2;
  - DISTINCT for unique Level 1 categories per Level 2 category
- List of broader categories that are popular according to user logs

Result - Level 2 categories

Top 10 unique Level 2 categories	
Category	**Frequency**
Film_naar_regisseur	941
Muziekalbum_naar_artiest	596
Burgemeester_van_een_voormalige_Nederlandse_gemeente	584
Nummer_naar_artiest	513
Plaats_in_India	399
Nederlands_burgemeester	369
Sport_naar_Nederlandse_gemeente	357
Gemeente_in_Noordrijn-Westfalen	315
Bouwwerk_naar_Nederlandse_gemeente	277
Rijksmonument_naar_gemeente	266

Method - Survey

- Target population:
  - Current users of DBpedia-NL
    - Difficult because finding out who users are is main problem
    - Published on DBpedia-NL website on May 18th, 2018
    - Spread over Twitter by DBpedia-NL chapter members
- Type of survey:
  - Self-administered questionnaire
  - Google Forms
  - No instructions given
  - 4 parts:
    - Respondent information
    - How do they use DBpedia-NL
    - What do they use DBpedia-NL for
    - Improvements

Method - Survey result analysis

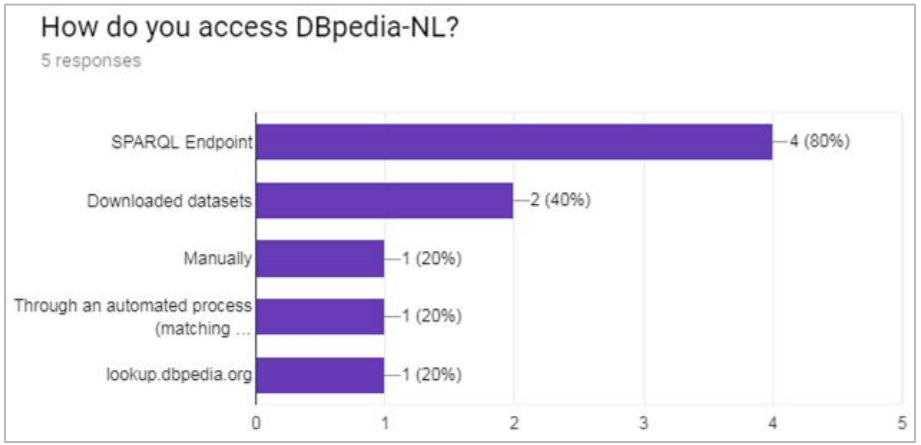
- Respondent information
  - Sort of respondents
  - Whether they use DBpedia-NL or not
- How do they use DBpedia-NL
  - Usage frequency
  - Way of accessing DBpedia-NL
  - Validation of data-driven method
- What do they use DBpedia-NL for
  - Categories
  - Validation of data-driven method
- Improvements
  - Current issues
  - Satisfaction

Result - Survey: Respondent information

- 5 respondents
- Aged 22 years or older.
- Fields of work:
  - Information architecture
  - Information Technology Development
  - Data science and heritage
  - Semantic Technology Provider
  - Software Engineer (in cultural heritage)

Result - Survey: How do they use DBpedia-NL

- 40% uses downloaded datasets
  - Does not show up in user logs



Result - Survey: What do they use DBpedia-NL for

- Survey results:
  - Geography
  - Books / Writers
  - Sport
  - Movies
  - Music
- Data-driven results:
  - Dutch & International municipality information
  - Books
  - Sport
  - Domestic and international movies
  - Music

Discussion

- Unused data:
  - Unused data Level 1 categories:
    - 787,039 unique URI requests in user log
    - 33.79% (265,926 URI requests) unused
  - Unused data Level 2 categories
    - 70,892 unique Level 1 categories
    - 6.72% (4,766 Level 1 categories) unused
- Data-driven method skewed by bots
- Validation through survey:
  - More time
  - Visit conferences

Conclusion

- Research questions:
  - What is a good way of characterizing user groups that use DBpedia-NL?
    - Proposed method is useful, but hard to validate through survey
  - What is DBPedia-NL currently used for?
    - Categories:
      - Domestic & International movies
      - Music
      - Sports
      - Dutch & International municipality information
      - Books
- High amount of bots

Questions?