Characterizing user groups of DBpedia-NL through user log analysis

FRANK WALRAVEN, Vrije Universiteit Amsterdam, Netherlands

Characterizing users of an online Linked Dataset is a difficult task due to the limited amount of information that is available. This paper proposes a data-driven method through user log analysis to characterize user groups that use the Dutch version of DBpedia and to determine what DBpedia-NL is currently used for. Knowing who its users are is important information to organisations such as DBpedia, since it allows them to focus the addition of new data in areas that their users are interested in. The described method uses subjects and hierarchical relations to characterize the log entries into grouped categories. These grouped categories are used to determine what subjects are popular on DBpedia and are validated using a survey. The popular categories that resulted from this method are Domestic & International Movies, Music, Sports, Dutch municipality information and Books.

1 INTRODUCTION

1.1 What is DBpedia?

DBpedia is one of the first and most prominent nodes of the linked open data cloud and holds a large-scale, multilingual knowledge base by extracting structured data from Wikipedia editions in 111 languages[14, 16]. The DBpedia Association was founded in 2014 to support DBpedia and the DBpedia Community. It is currently situated in Leipzig, Germany and affiliated with the non-profit organisation Institute for Applied Informatics (InfAI)[7]. Auer and Bizer[9] described it as "a community effort to extract structured information from Wikipedia and to make this information available on the web".

DBpedia is freely available to anyone. According to the official DBpedia Wiki[2] it currently has thousands of users such as large companies, libraries, researchers and web developers. Due to the massive size of this project DBpedia is divided into local chapters that coordinate their respective language on DBpedia. These local chapters are coordinated by the DBpedia Internationalization Committee. This research focuses on the Dutch chapter of DBpedia, called DBpedia-NL, but the proposed methods can be used on the international version of DBpedia as well. DBpedia-NL maintains and expands the Dutch side of DBpedia data.

All releases of the DBpedia knowledge base can be downloaded[6] and 14 out of the 111 language editions can be accessed using SPARQL queries[16]. An analysis of the period between October 2016 and December 2017 shows that the DBpedia service had on average 7,343,939 hits per day[5]. Clearly the DBpedia service is highly used, which makes it very important that this service is constantly being improved and expanded.

There are several ways in which anyone can help improve DBpedia. Some ways of contributing to DBpedia are given on the website[8], such as joining Community Meetings or the Ontology Committee, providing answers to user questions and giving feedback through the bugtracker[1].

At the moment it is unclear to the DBpedia Association in general and to DBpedia-NL specifically who its users are exactly and what their expectation of the Linked Dataset is. According to the members of the Dutch DBpedia chapter this information is much needed in order to shape DBPedia-NL and it's extensions in a way that caters to its current users. For example, if the DBpedia organisation knows that a large part of their users are interested in very specific information about insects then that knowledge can be used to improve and better market DBpedia through adding more data on that subject. Linked Datasets such as DBpedia require a lot of effort to expand, which means that knowing in what directions your users want you to go is of great value.

1.2 Research questions & hypothesis

The DBpedia Association and specifically DBpedia-NL wants to improve its dataset. In order to do this they need to know who its current users are and what they currently use it for. Based on that information DBpedia-NL can then focus on the areas that its users are interested in when improving their dataset. This makes

Author's address: Frank Walraven, Vrije Universiteit Amsterdam, Netherlands, frankwalraven@gmail.com.

[1]DBpedia bugtracker: https://github.com/dbpedia/extraction-framework/issues

for a useful case study for a data-driven method based on analysizing the user logs that can be expanded to other Linked Datasets as well. The proposed data-driven method will be used to characterize the user groups of DBpedia-NL through its popular categories. Based on those results a determination is made on what DBpedia-NL is currently used for. The research questions this paper aims to answer are:

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

Based on the official DBpedia Wiki[2] we hypothesise that the current users of DBpedia-NL are mostly linked data researchers, libraries, web developers and large companies. However, as was mentioned before the DBpedia Association itself currently does not know who their users are exactly, so it is highly possible that the resulting user groups include unexpected results.

1.3 Related Work

Using user logs for analyzing databases and websites has been looked at by several researcher already, such as Wang, Berry, & Yang[19] and Jansen, B. J.[15]. Such logs are described by Jansen, B. J. as "an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine". Analyzing user logs to obtain information is called Transaction Log Analysis (TLA). Davis, P. M.[11] defined Transaction Log analysis as "a non-intrusive method for collecting data from a large number of individuals for the purpose of understanding online-user behavior". TLA consists of three stages: collection, preparation and analysis (Jansen, B. J.[15]. The collection stage is the process of collecting the relevant data for a defined period. After this collection of data it has to be pre-processed to be usable for analysis. These three steps will be followed in this paper's proposed method as well.

1.3.1 User log extraction and analysis.

The characterization of user groups can be done in several ways. One possible way is the use of machine learning, namely data mining. Hand, D. J.[13] defined data mining as the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Regression, classification and clustering are the data mining techniques that are mainly used with the clustering technique being the best choice for very large datasets (Sharma and Bajpai[18]). Berkhin[10] describes clustering from both a general and a machine learning perspective. Clustering in general is described as "a division of data into groups of similar objects" and clustering from a machine learning perspective is described as "[...] clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept". Sharma and Bajpai[18] describe cluster analysis as "Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The greater the likeness (or homogeneity) within a group, and the greater the disparity between groups, the better or more distinct the clustering." The method that is proposed in this paper follows the general perspective of clustering; The user log entries are grouped into clusters using their subjects and hierarchical relations.

At the moment of writing this paper there is not much information available on the users of DBpedia. According to the official DBpedia wiki[2] it currently has thousands of users such as large companies, libraries, researchers and web developers.

1.3.2 Query analysis.

Harry Halpin[12] aimed to answer whether there is anything worth finding for ordinary users in Linked Data. This was done through inspecting what information needs users are expressing using a hypertext search engine and then using a sample of this data to test if Linked Data can satisfy these information needs. Limam L and Coqul D[17] did a similar research in which they propose an enhancement of search query log analysis by taking into account the semantic properties of query terms. They defined a query terms clustering algorithm that is used to extract user interests in the following way: "an algorithm based on the semantic distance [...] we define a cluster as a set of query terms such that the distance between each pair of query terms of the

set is inferior to a predefined threshold." Even though their methods proved useful it is not necessary to group our user log entries on their semantic distance to create clusters because we use a Linked Dataset which already has relations build into it that can be used to create such clusters.

2 METHODOLOGY

This method proposed in this paper consists of two parts; A data-driven method and a survey. The data-driven method uses the subjects and hierarchical relations to characterize the log entries into grouped categories. The resulting list of popular categories is then validated using the survey.

2.1 Used software and data

In this paper a combination of the software is used depending on the needed functionality and possible computational restrictions. All used software can be found in Table 1.

Table 1. Used tools

Task	Software
Pre-processing data	Notepad++[1]
SQL querying	Office Access 2016[2]
Analyzing data	Office Excel 2016[2]
Survey	Google Forms[3]

The data that will be analyzed are the user logs and a local dump of DBpedia-NL[4]. The user logs of DBpedia-NL that are used in this paper span the 6th of December 2017 until the 14th of January 2018 and consists of 4,426,543 entries. These user logs contain sensitive information of DBpedia-NL's users, which is why all IP-addresses are anonymized. To obtain these user logs you have to contact the DBpedia Association. Example files of the used data can be found in Appendix 13. Due to the large size of Office Access 2016 files when importing large datasets only an example file is available; All needed queries are given in this paper. An attempt was also made to query DBpedia-NL through the SPARQL endpoint without using local dumps, however due to the large size of both the user logs and DBpedia-NL this resulted in a runtime of approximately 21 days just to retrieve the subject relations of the log entries. A runtime that long is not feasible in the time that was set for this research, which is why a local dump of DBpedia-NL is used and highly recommended.

2.2 Pre-processing data

2.2.1 *User logs.*

The user logs are made up of 4,426,543 entries, of which an anonymized example line can be seen below:

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

Office Access 2016 is used to query the local DBpedia-NL dump and the user log entries. Office Access 2016 tends to give errors when the data it uses includes apostrophes and brackets, which is why a pre-processing step is done using Notepad++. All characters that are unneeded and that Office Access 2016 has issue with are removed: apostrophes, brackets, minus, "GET " and the HTTP information. This results in the following pre-processed entry:

```
xx.xxx.xxx.xxx 07/Dec/2017:03:34:23 +0100
/resource/Roger_Cicero
```

This pre-processed user log is then imported into Office Access 2016 where it can be queried on its hierarchical relations.

2.3 Data analysis

2.3.1 *IP-addresses in the user logs.*

To get a better sense of the data that can be found in the user log a first look is taken at what kind of users are behind each URI request, if that information is available. This is done through analysizing the IP-addresses that can be found in the user logs. IP-addresses can be divided into several classes[4] based on the range of their first octet. The first octet is the part of the IP-address before the first period. For example, the first octet of IP-address 138.100.200.30 is 138. The exact range of each class can be found in Table 2.

[1]Notepad++: https://notepad-plus-plus.org/download/v7.5.6.html
[2]Office download: https://products.office.com
[3]Google Forms: https://docs.google.com/forms
[4]DBpedia-NL dump: http://downloads.dbpedia.org/current/core-i18n/nl/

Table 2. IP-address classes[4]

IP-address classes		
Class	**Start address**	**Finish address**
A	0.0.0.0	126.255.255.255
B	128.0.0.0	191.255.255.255
C	192.0.0.0	223.255.255.255
D	224.0.0.0	239.255.255.255
E	240.0.0.0	255.255.255.255

Using the class of each IP-address it is possible to extract what kind of network is behind this IP-address[4]. These class descriptions can be found in Table 3.

Table 3. IP-address class description[4]

Class descriptions	
Class	**Description**
A	Very large networks such as multinational companies
B	Large networks such as a college and Internet Service Providers
C	Small to mid-sized companies
D	Multicast services
E	Reserved for experimental use

These class description are more of a general description for what kind of network and / or user is behind an IP-address, but are not conclusive. Aside from analyzing the IP-address classes an IP-address lookup[1] is also done on the most used IP-addresses to get a better understanding of the kind of user that is behind those specific IP-addresses. If available, the relevant information that is returned is the Host, Country, IP owner info and Domain owner info. It is important to note that not all IP-addresses have all that information available, which means that some IP-addressess can not be classified this way. One or all of these can hold information that shows whether an IP-address belongs to a bot, a specific organisation or otherwise. For example, part of the returned info of a specific IP-address looks like this:

```
IP: 68.180.228.46
Host: b110018.yse.yahoo.net
Country: United States
```

In this example the Host information tells us that this IP-address belongs to Yahoo. If this same IP-address made thousands of URI requests that would likely make it a bot, since Yahoo is a search engine that uses web crawlers. Sometimes the Host info is not as clear, in which case an in-depth look is taken at the IP owner info and the Domain owner info. The IP-addresses that can not be identified using an IP-address lookup will be classified as "unsure".

It is also possible to see whether different IP-addresses are coming from the same larger network. If the first and second octet of different IP-addresses are the same then that means that they come from the same larger network.

2.3.2 URI request distribution.

Aside from analyzing who the users are through their IP-addresses the distribution of the URI's that have been requested are looked at to get a better sense of the data that is used in this paper. The distribution of these URI Requests is expected to be exponentially distributed due to bots also being apart of the user logs, which are likely to have make a large amount of requests to DBpedia.

2.3.3 URI categories.

One of the goals of this research is to characterize who the user groups that use DBpedia-NL are. An attempt to answer this research question is made through analyzing the user logs and their categories. These categories are created using the log entries' corresponding subject and hierarchical relations. Based on these popular categories we can get a sense of what kind of users there are on DBpedia-NL.

2.3.3.1 Level 1.

Wikipedia Categories are represented in DBpedia using the SKOS vocabulary and DCMI terms[3]. DBpedia-NL resources can have a "dcterms:subject" property which relates that specific resource to a corresponding category. For example, the "Android_TV" resource has a dcterms:subject property that relates to "category-en:Google" and
"category-en:Android_(operating_system)_software".
These related categories will be henceforth referred to as Level 1 categories.

Using a combination of the local DBpedia-NL dumps, the user log entries and SQL all the Level 1 categories are extracted. This results in a list of URI Requests and their corresponding Level 1 categories, which can be sorted on the amount of times each Level 1 category is found in the user log. This is done using the SQL

[1] IP-address lookup - https://ip-lookup.net

query below in Office Access 2016 and then sorted using the PowerPivot function of Office Excel 2016. All SQL queries that are shown are in pseudo-code to ensure clarity.

```
SELECT level_1_categories.categories,
uri_list.uri
FROM level_1_categories INNER JOIN uri_list
ON level_1_categories.uri = uri_list.uri;
```

The user log includes several specific URI request that are made multiple times by the same user, which will skew the results towards those URI requests. To ensure that the results are as clear and valid as possible the frequency of unique URI's per Level 1 category is calculated using the same SQL query in combination with DISTINCT.

The normalized frequencies are also calculated to compensate for resources with high in-degree. This is done because a Level 1 category such as "American movies" has a larger degree than Level 1 category "Komische films". This results in a very large Level 1 category having a larger frequency of found resources in the user log purely because that specific category is related to a much larger amount of resources. To check how reliable the results are the frequency of the Level 1 category results are normalized using the following equation:

$$NormalizedFrequency = \frac{\#\ of\ UNIQUE\ URI's\ in\ user\ log\ per\ Level\ 1\ Category}{Total\ \#\ of\ linked\ URI's\ in\ DBpedia-NL\ per\ Level\ 1\ Category} * 100$$

This created list of Level 1 categories and their corresponding URI requests will then be queried in combination with the user logs to get the amount of non-unique and unique ip-addresses that looked for each Level 1 category. In order to get the amount of non-unique ip-addresses per Level 1 category the following semi-formal SQL query is used:

```
SELECT COUNT(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;
```

Normally a query using both COUNT and DISTINCT is written in the following format:
"SELECT count(DISTINCT ..) FROM .. ".
However, Office Access 2016 uses the Access-Engine, which does not support the usage of COUNT and DISTINCT in this format. To do this in Office Access 2016 the query has to be formatted like this:

```
SELECT count(*)
FROM
(SELECT DISTINCT Name FROM table)
```

Due to this limitation the SQL query that is used for the amount of unique IP-addresses per level 1 category is formatted slightly differently:

```
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;
```

Lastly, the distribution of the frequency of different IP-addresses per Level 1 category is analyzed to determine how evenly distributed the usage of DBpedia-NL is within its users.

2.3.3.2 Level 2.
Level 1 categories can also have a "skos:broader" relation, which is used to connect subcategories and supercategories. In this paper it relates Level 1 categories to a corresponding broader category. These broader categories will be henceforth referred to as Level 2 categories. Each Level 1 category is queried on their "skos:broader" relationship using the local DBpedia-NL dumps and SQL queries. For each Level 2 category both the frequency of Level 1 categories and the frequency of unique Level 1 categories per Level 2 category are

gathered using the SQL query below in Office Access 2016 and Office Excel 2016's Powerpivot functions. Note that DISTINCT should be added to get the unique Level 1 categories.

```
SELECT level_1_category_list.level1cat,
dbpedia_skos_broader.level2cat
FROM level_1_category_list
INNER JOIN dbpedia_skos_broader
ON level_1_category_list.level1cat =
dbpedia_skos_broader.level1cat;
```

This query results in a list similar to the previously described Level 1 categories, but instead it allows us to use Office Excel 2016's PowerPivot function to see which broader categories are mainly searched for in the user logs by looking at the different frequencies of each Level 2 category.

2.3.3.3 Unused data.
It is important to note that querying the user logs on their "dcterms:subject" and "skos:broader" relationship will result in some of the user log data being unused, since some resources do not have such a relationship. The amount of unused data is reported in the Results section and discussed in the Discussion section.

2.4 Survey

Since the previously described data-driven method uses the user log entries as its data there is a possibility that parts of the user log entries are made up of bots such as webcrawlers that search engines use. This results in a list of popular categories amongst both bots and human users, but we are mainly interested in the human user interests. In order to verify the results of the data-driven method a survey is conducted that is sent to human users. The popular categories that follow from this survey are then compared to the results of the data-driven method to determine whether or not the possible bots in the data skew the results.

2.4.1 Survey target population.

The respondents this survey aims to reach are current users of DBpedia-NL, which will be a hard task since one of the goals of this research is to find out who DBpedia-NL's current users actually are. The survey has been published on the official DBpedia-NL website on May 18th, 2018[1]. It is expected that most of the active users of DBpedia-NL visit this website, which would result in respondents that fit its user base. Aside from publishing on the official website the survey has also been distributed over Twitter by members of the DBpedia-NL chapter as well as sent to the DBpedia-NL mailing list[2]. It is important to note that publishing the survey on the official DBpedia-NL website likely only results in respondents that are among the most active users of DBpedia-NL, because those will be regularly visiting that website. This is an important group that has to be included in the survey results, however users that only use DBpedia a few times a year are not expected to visit that website frequently. Following that expectation that means that this less active group is likely not going to find this survey. This could skew the survey data towards only very active DBpedia-NL users, which is why the survey is also shared through Twitter by the DBpedia-NL chapter members to try to counteract this possible issue.

2.4.2 Type of survey.

This survey is designed by the author of this paper as a self-administered questionnaire created on Google Forms and there are no instructions given to the respondents other than a request to fill in the survey. The survey questions, question type and question division can be found in Table 4 and the full survey can be found on the Google Forms page[3]. The survey questions are divided into 4 parts:

(1) Respondent information
(2) How do they use DBpedia-NL
(3) What do they use DBpedia-NL for
(4) Improvements

[2]DBpedia-NL mailing list - dbpedia-dutch@lists.sourceforge.net
[3]Google Forms Survey - https://goo.gl/forms/vHvSD7FY86Emw1qI3

Table 4. Types of questions in the survey

Type of questions in survey		
Question	**Type of question**	**Part**
Have you ever used DBpedia?	Discrete	Respondent information
Have you ever used the Dutch version of DBpedia (DBpedia-NL)?	Discrete	Respondent information
What is your field of work?	Open	Respondent information
How often do you use DBpedia-NL on average?	Closed (multiple choice)	How do they use DBpedia-NL
How do you access DBpedia-NL?	Closed (checkboxes)	How do they use DBpedia-NL
What do you use DBpedia-NL for?	Closed with "Other" option (checkboxes)	What do they use DBpedia-NL for
In your own words, how do you use DBpedia-NL?	Open	How do they use DBpedia-NL
What specific categories of information do you mainly use DBpedia-NL for?	Closed with "Other" option (checkboxes)	What do they use DBpedia-NL for
What specific components of DBpedia-NL do you think need to be expanded on?	Open	Improvements
How satisfied are you with the current state of DBpedia-NL?	5-point Likert scale	Improvements
Could you explain why you are (un)satisfied?	Open	Improvements
What is your age?	Closed (multiple choice)	Respondent information

2.4.3 *Survey result analysis.*

Because the survey is divided in four parts this is used to analyze its results. The Respondent information part is used to determine whether the respondent actually uses DBpedia-NL and what kind of work they do. The second part of questions ("How do they use DBpedia-NL") is used to determine the usage frequency and way of accessing DBpedia-NL. The third part ("What do they use DBpedia-NL for") goes more in-depth on what exactly DBpedia-NL is used for by the respondent. The fourth and final part, Improvement, is used to get a sense of which issues DBpedia-NL currently has.

First, all respondents that have never used DBpedia-NL are filtered out of the results to ensure that all data is usable to validate the described data-driven method. The "Respondent Information" and "What do they use DBpedia-NL for" parts are compared to the results from the data-driven method that is conducted in this research. For example, if the data-driven results tell us that most users of DBpedia-NL are interested in "municipality information" then this can be compared to the "Respondent Information" results of the survey. If these results match it can be concluded with a high amount of certainty that that is the correct result.

The "How do they use DBpedia-NL" part of the survey is used to get more information regarding the way users access DBpedia-NL. It is possible that a large part of DBpedia-NL's users only uses a downloaded version of the DBpedia-NL database. The data-driven method uses the DBpedia-NL user logs as its basis, meaning that it does not include users that did not use the SPARQL Endpoint. If this is the case for a significant part of the survey respondents that would also mean that a large part of DBpedia-NL's users can not be found in the results of the data-driven method.

Lastly, the "Improvements" questions were added to get an idea of how satisfied the users are with DBpedia-NL and what they are currently missing. This is not used to check the results of the data-driven part of this research, but could be useful for DBpedia-NL regardless.

3 RESULTS

3.1 IP-addresses analysis

Table 14 in the Appendix shows the top 30 most used IP-addresses, their corresponding IP-class and whether or not it is a bot. Whois information does not always explicitly show what organization an IP-address belongs to, but if this is the case the corresponding organization is shown as well. For example, the IP-address with the highest frequency is 68.180.xxx.xx. An IP-address lookup as was described in the methodology returns the following basic information:

```
IP: 68.180.xxx.xx
Host: b110018.yse.yahoo.net
Country: United States
```

This example shows that this IP-address belongs to Yahoo. In some cases the organization is not shown so clearly in the Hostname, in which case the detailed IP owner info that is given when executing an IP-address lookup could still hold such information. An IP-address is classified as a bot when this is shown in its Whois or Host information or in the case of the previous example if they are coming from a search engine. Note that some of these IP-addresses did not have any WHOIS information attached to them when doing an IP-lookup, in which case they are classified as "unsure".

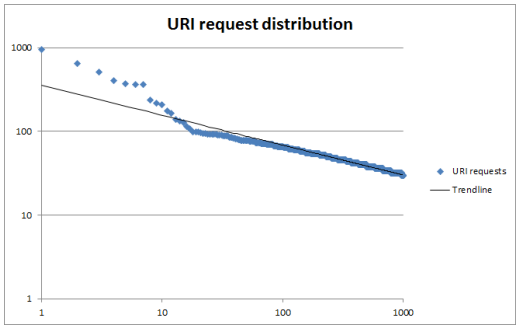
If an IP-address is within the same larger network as another IP-address in this list then they are given the same background color in the table. If the background color is white that means that this IP-address is the only one in its larger network that is found in this list.

80% of the IP-addresses in Table 14 falls under Class A and is also classified as a bot. To get a more in-depth look into the IP-addresses found in the user log all Class A IP-addresses are excluded resulting in Table 15 in the Appendix. Even after removing all Class A IP-addresses and using their WHOIS and host information to identify whether or not they are bots 60% of the most used IP-addresses are still classified as bots, while only 10% are not. The remaining 30% can not be classified as a bot or not. This suggests that most of DBpedia-NL's bulk consumers are bots.

3.2 URI request distribution

Figure 1 shows that the logarithmic distribution of the URI requests follows a fairly straight line. This suggests that the URI requests follow an exponential distribution. This is in line with the user logs being skewed due to a large amount of bots, since those bots are likely to make a large amount of URI requests.

Fig. 1. Log-log graph of the frequency of each URI request using a random sample of 100.000 URI requests



3.3 Level 1 categories

Not all resources have a corresponding Level 1 category, which means that some of the data within our dataset gets lost using this method. 787,039 unique URI's can be found in the user log, of which 33.79% (265,926 URI's) has no Level 1 category.

Table 5 shows the Level 1 categories with the highest frequency sorted in descending order.

Table 5. Top 10 level 1 categories

Top 10 Level 1 categories (non-unique uri's per Level 1 category)		
Level 1 category		**Frequency**
1	Amerikaanse_film	9001
2	IUCN-status_niet_bedreigd	5591
3	Dramafilm	4629
4	Nederlands_voetballer	3707
5	Amerikaans_acteur	3665
6	Amerikaans_filmacteur	3302
7	Amerikaans_televisieacteur	2879
8	Komische_film	2534
9	Pseudoniem	2380
10	Dier_uit_het_Palearctisch_gebied	2376

The methodology explained that it is possible that a specific URI request is made a lot of times, which will skew the resulting data. Table 6 contains the frequency of unique URI requests that have been made per Level 1 category, as well as their corresponding normalized frequency percentages.

Table 6. Top 10 distinct level 1 categories

Top 10 distinct Level 1 categories			
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%

When just looking at the amount of unique URI's per Level 1 category the top 10 does not change much. Only number 4 and 5 switch spots. However, the normalized frequency results tell us how big each Level 1 category count actually is within DBpedia-NL. The "IUCstatus_niet_bedreigd" category is number 2 in the top 10 most unique URI's per Level 1 category, but that still is only 27% of the total amount of resources that are linked to this Level 1 category on DBpedia-NL. This means that because this Level 1 category contains so many resources we can not say that this is a popular Level 1 category per se. The same goes for the "Dier_uit_het_Palearctisch_gebied" category.

3.3.1 *IP-addresses per Level 1 category.*

Each URI request has a corresponding IP-address, which are used to determine how many IP-addresses in total fall within each Level 1 category. Table 7 contains the frequency of all IP-addresses per category.

Table 7. Top 10 Level 1 categories with the highest frequency of IP-addresses

Top 10 Level 1 categories with most non-unique IP-addresses	
IP frequency	**Level 1 category**
89,778	Amerikaanse_film
46,456	Dramafilm
44,266	IUCN-status_niet_bedreigd
41,902	Amerikaans_acteur
38,874	Amerikaans_filmacteur
37,750	Nederlands_voetballer
30,760	Pseudoniem
30,016	Amerikaans_televisieacteur
29,932	Land
25,388	Komische_film

To get a better understanding of how many different users made URI requests falling within the Level 1 category the number of unique IP-addresses per Level 1 category can be found in Table 8.

Table 8. Top 10 Level 1 categories with the highest frequency of distinct IP-addresses

Top 10 Level 1 categories with most unique ip-addresses	
Unique IP frequency	**Level 1 category**
596	Nederlands_voetballer
431	Amerikaanse_film
408	Amerikaans_acteur
389	Pseudoniem
381	Amerikaans_filmacteur
365	IUCN-status_niet_bedreigd
342	Dramafilm
341	Belgisch_voetballer
335	Amerikaans_televisieacteur
333	Plaats_in_Gelderland

Note that both “Land” and “Komische_film” have been replaced by “Belgisch_voetballer” and “Plaats_in_Gelderland”. Table 8 also shows that “Nederlands_voetballer” is in actuality the most used Level 1 category by different users, which is more in line with the expectation that DBpedia-NL would be used mainly for Dutch subjects.

Fig. 2. Log-log distribution of the frequency of different IP-addresses within Level 1 categories

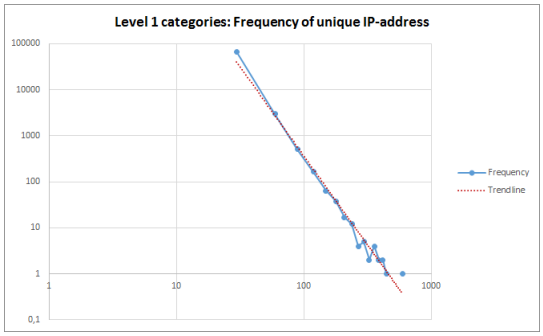


Figure 2 shows the log-log distribution of the frequency of different IP-addresses within Level 1 categories. The corresponding data can be found in Table 9.

Table 9. Frequency of distinct IP-address ranges within Level 1 categories

Frequency of unique IP-address ranges within Level 1 categories	
Upper limits	**Frequency**
29,75	66.750
59,5	2.995
89,25	505
119	164
148,75	64
178,5	38
208,25	17
238	12
267,75	4
297,5	5
327,25	2
357	4
386,75	2
416,5	2
446,25	1
476	0
505,75	0
535,5	0
565,25	0
596	1

66,750 different IP-addresses have made between 1 and 30 URI requests that fall within a Level 1 category. The graph shows a steep drop and has a straight trendline, meaning that it is exponentially distributed. This means that most users made at most 30 different URI requests.

Due to such a large amount of users falling within this 1 to 30 URI requests range an in-depth analysis of the distribution of this group is done, which can be seen

in Figure 3. The corresponding data can be found in Table 10.

Fig. 3. Log-log distribution of the frequency range 1 to 31 of different IP-addresses within Level 1 categories

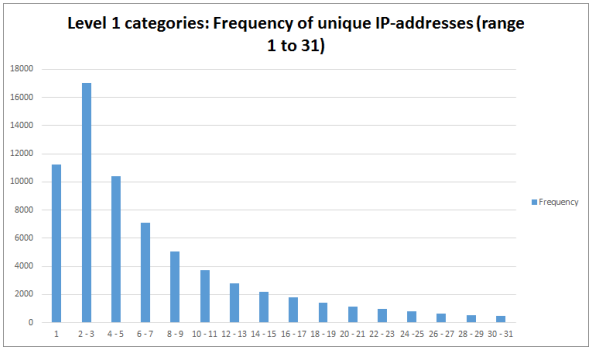


Table 10. Frequency of unique IP-address ranges within Level 1 categories (in-depth look at the largest frequency group: 1 - 31)

Frequency of unique IP-address ranges within Level 1 categories (in-depth look at largest frequency group: 1 - 31)	
Upper limits	**Frequency**
1	11.239
3	17.008
5	10.410
7	7.068
9	5.052
11	3.707
13	2.808
15	2.175
17	1.814
19	1.386
21	1.125
23	947
25	812
27	659
29	540
31	463

This distribution shows that most users made either 2 or 3 different URI requests that fall within the Level 1 category.

3.4 Level 2 categories

Similar to the Level 1 category results the Level 2 categories also have some unused data. This is due to not all Level 1 categories having a related Level 2 category. The dataset holds 70,892 unique Level 1 categories, of which 6.72% (4,766 Level 1 categories) has no Level 2 category.

Table 11 shows the Level 2 categories with the highest frequency sorted in descending order.

Table 11. Top 10 Level 2 categories

Top 10 Level 2 categories	
Category	**Frequency**
Film_naar_genre	19,680
Voetballer_naar_nationaliteit	17,204
Film_naar_land	14,700
Cinema_in_de_Verenigde_Staten	14,417
Film_naar_jaar	13,854
Plaats_in_de_Verenigde_Staten	11,671
Nederlands_sporter	8,982
Soort_naar_IUCN-status	8,698
Olympisch_deelnemer_naar_nationaliteit	7,685
Film_naar_regisseur	7,417

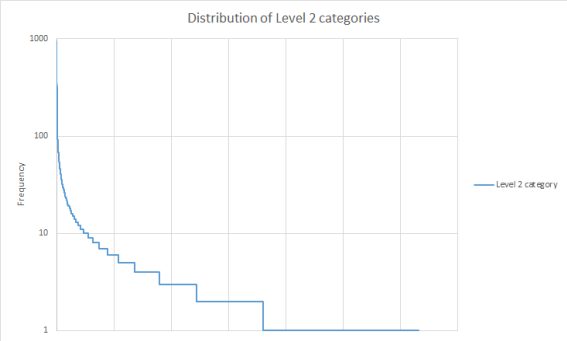
Similar to the previous Level 1 categories section it is possible that specific Level 1 categories show up a lot, which risks skewing the resulting data. Table 12 contains the frequency of unique Level 1 categories per Level 2 category.

Table 12. Top 20 distinct Level 2 categories

Top 10 distinct 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
Compositie_naar_componist	254
Kerkgebouw_naar_Nederlandse_gemeente	235
Spanner	195
Boek_naar_schrijver	193
Compositie_naar_jaar	185
Hoofdstad	183
Voetballer_naar_nationaliteit	180
Geografie_naar_Nederlandse_gemeente	177
Lijsten_van_voetbalinterlands	176
Heer_of_vrouwe	160

Figure 4 shows the log-log distribution of the frequency of unique Level 1 categories per Level 2 category. Only a relative small group of Level 2 categories has a high frequency, with the average frequency of unique Level 1 categories per Level 2 category being 4.

Fig. 4. Log-log distribution of the frequency of different IP-addresses within Level 1 categories per Level 2 category



3.5 Survey

All graphs that are referenced in this section can be found in the Appendix.

3.5.1 *Amount of respondents.*

As was mentioned in the methodology it is to be expected that finding respondents for this survey is difficult, because at the time of doing this research determining who DBpedia-NL's users are is the main question to answer. Even though this survey was shared in multiple ways it still got only 5 respondents. This means that validating the results that we got from this survey with our data-driven results will not give a conclusive answer on whether or not those data-driven results are correct. However, it does give a very small glimpse into DBpedia-NL's current users.

3.5.2 *Respondent information.*

All respondents have used DBpedia as well as DBpedia-NL and are aged 22 years or older. These respondents work in the following fields:

(1) Information architecture.
(2) Information Technology Development.
(3) Data science and heritage.
(4) Semantic Technology Provider.
(5) Software Engineer (in cultural heritage).

3.5.3 *How do they use DBpedia-NL.*

As can be seen in Figure 5 two of the respondents use DBpedia-NL on average every day, one uses it twice a week and the other two respondents use it once in 6 months and once in a year.

Two out of the five respondents use a downloaded version of DBpedia-NL (Figure 6). This means that 40% of the respondents to this survey have used DBpedia-NL without it being logged in the user logs. This suggests that the user log is more skewed than previously thought; It is not just made up of a large amount of bots, it also is missing part of the DBpedia-NL usage that was done through local DBpedia-NL dumps.

3.5.4 *What do they use DBpedia-NL for.*

Figure 7 shows that three of the 5 respondents use DBpedia-NL for research and business solution purposes, with four of them using it for personal use as well.

When asked to describe in their own words what they use DBpedia-NL for the following responses were given:

(1) Denote common meanings for resources in organization-bound namespaces
(2) Fact checking and information retrieval
(3) Get structured data from Wikipedia.
(4) Seeking enrichment for resources in other datasets by linking those resources to nl.dbpedia. Also for personal knowledge acquisition.
(5) I use DBPedia lookup to annotate web resources (not 100% sure if this also accesses what you call DBpedia-NL). Currently the link with DBpedia is only for demonstration purposes

The main goal of this survey was to compare the popular categories that come out of the survey results with the popular categories that were gotten through the data-driven method. Due to the small amount of respondents this is not viable, however this is still done to ensure completeness.

The categories that were most popular (Figure 8) with 3 out of 4 respondents according to the survey results are Geography and Books / Writers. Sport, Movies and Music are the second popular categories with 2 out of 5 respondents chosing those categories. The data-driven method resulted in the following categories being the most popular: Domestic and international movies, Music, Sports, Dutch municipality information and International municipality information.

This partly fits the results of this survey, since Books / Writers is shown to be the most popular category

in the survey and is in the list of most popular categories of the data-driven method as well. Geography can not be found in the list of data-driven method results. The second most popular categories according to the survey results are Sport, Movies and Music. These categories were also shown to be popular in the data-driven method results. This means that even though the amount of respondents is very low, the survey results support the results of the data-driven method.

3.5.5 *Improvements.*

Lastly, the survey asked about what could be improved upon on DBpedia-NL. These were open questions of which the full answers can be found in Table 16 . Summarized, these answers talk about needing better and more complete data quality with more links to other sources. Another recurring answer is that DBpedia-NL has to be integrated with Wikidata. The satisfaction of the respondents was rated on a 5-point likert scale from very unsatisfied to very satisfied and can be found in Figure 9. Two out of 5 respondents rated it at 4 with the rest of the respondents rating it at 1, 3 or 5 respectively.

3.6 Results of combined data

Based on the results of the data-driven method and the survey we conclude that the bulk users of DBpedia-NL are bots, which at least partly consists of search engine crawlers. The URI requests in the user logs are exponentially distributed, which means that a small amount of resources is requested a very large amount of times. This is likely also due to the high amount of bots that can be found in the user logs. Another goal of this paper was to identify what DBpedia-NL is currently used for. We used the previously described Level 1 and Level 2 categories to get a list of broad subjects that are most popular on DBpedia-NL. The most popular subjects are:

- Domestic and international movies
- Music
- Sports
- Dutch municipality information
- International municipality information
- Books

These results are supported by the survey results, however the survey had too few respondents to be considered a valid test.

4 DISCUSSION

There are a few important notes to make regarding this paper. As was described in the methodology some data from the user logs is unused when employing the described data-driven method due to those resources not having a related Level 1 category. In this particular user log this meant that 33.79% of the data was unused in the Level 1 category step and 6.72% of that was unused in the Level 2 category step. This means that a significant part of the user logs does not show up in the results of this paper at all and could result in some popular categories not being found. On the other hand, the user log consists of 4.426.543 entries, which even with part of the data being unused is large enough to get a good sense what the popular categories on DBpedia-NL are.

Another important note to make is that the results show that a large part of the users in this user log consists of bots. The resulting list of popular subjects on DBpedia-NL is heavily influenced by these bots and does not necessarily give a conclusive answer on the question of which specific categories are popular with human users. For example, the results show that Dutch municipality information is a popular category on DBpedia-NL. This could be because the search engine crawlers use DBpedia-NL to lookup such information constantly to make sure their search engine stays up-to-date. That does not mean that that category is actually often looked for on DBpedia-NL directly. The resulting categories from the data-driven method are also very spread out, which is likely the result of the high amount of bots that can be found in the user logs.

Even though the data-driven method that is proposed in this paper is not able to strictly determine who the users are of DBpedia-NL, it still proves useful in determining the interests of the users of a Linked Dataset based on the user logs. For example, knowing which categories are popular and that there is a high amount of bots within its user groups is useful information when determining how to expand and improve your Linked Dataset. When using the proposed methods on a different Linked Dataset we recommend the researcher puts more time into getting enough respondents to their survey in order to make sure you can properly validate your data-driven results. Reaching respondents of a Linked Dataset survey is clearly a difficult task, which

is why we suggest visiting conferences etc. where users of your specific Linked Dataset are expected to be and conduct the survey there.

Lastly, the goal of the survey was to use it as a test of the results that were gotten through the data-driven method. The survey results did support the results from the data-driven method, but the amount of respondents was too low to make this a valid test. If the survey had more respondents it would have told us which categories are popular among the human users, which would allow us to more conclusively know what kind of human users can be found on DBpedia-NL. When using the described methods in this paper it is recommended that the survey is conducted at local events where DBpedia-NL users can be found in order to get more respondents. This was not possible in this research due to limitations in time.

5 CONCLUSION

In this paper a data-driven approach is used to try to characterize the user groups that use DBpedia-NL. The main research question in this paper is "What is a good way of characterizing user groups that use DBpedia-NL", which we attempted to answer using a data-driven method that analyses the user logs of DBpedia-NL in combination with a local DBpedia-NL dump and a survey to validate the results of said data-driven method. The resulting popular categories based on both parts of the method are Domestic and International Movies, Music, Sports, Dutch municipality information, International municipality information and Books. It is also shown that a large part of the user log is made up of log entries coming from bots, meaning that these results are likely skewed by those bots and do not conclusively give an answer on which human users use DBpedia-NL.

The second research question of this paper is "What is DBpedia-NL currently used for?". The described data-driven method does answer this question conclusively; It is currently used to search for information on the previously mentioned most popular categories by bots and human users.

6 APPENDIX

Table 13. Download links to data

File description	**Download link**
PowerPivot Table - Resources per Level 1 category	https://figshare.com/s/a9fd218e014c3b8fb4ba
PowerPivot Table - Level 2 categories	https://figshare.com/s/0543194e41c32dcb42d4
Office Access 2016 example - Queries for resources and Level 1 categories	https://figshare.com/s/b57a2e272182c165dd59
DBpedia-NL dump	http://downloads.dbpedia.org/current/core-i18n/nl/

Table 14. IP-addresses that used DBpedia-NL (anonymous IP-addresses)

Table 3.1a - Ip-addresses that used DBpedia-NL				
#	**IP-address**	**Frequency**	**IP-class**	**Bot?**
1	68.180.xxx.xx	1,981,562	A	Yes (Yahoo)
2	17.142.xxx.xxx	782,012	A	Yes (Apple)
3	130.251.xx.xxx	130,780	B	Unsure
4	68.180.xxx.xx	123,090	A	Yes (Yahoo)
5	78.46.xxx.xxx	82,204	A	Unsure
6	46.229.xxx.xx	53,190	A	Yes
7	46.229.xxx.xx	51,658	A	Yes
8	46.229.xxx.xx	51,314	A	Yes
9	46.229.xxx.xx	51,286	A	Yes
10	46.229.xxx.xx	51,192	A	Yes
11	46.229.xxx.xx	51,164	A	Yes
12	46.229.xxx.xx	50,690	A	Yes
13	46.229.xxx.xx	50,656	A	Yes
14	46.229.xxx.xx	50,414	A	Yes
15	46.229.xxx.xx	50,326	A	Yes
16	46.229.xxx.xx	50,146	A	Yes
17	46.229.xxx.xx	44,992	A	Yes
18	46.229.xxx.xx	43,850	A	Yes
19	138.201.xxx.xx	41336	B	Unsure
20	46.229.xxx.xx	41,327	A	Yes
21	46.229.xxx.xx	40,930	A	Yes
22	202.180.xx.xxx	32,570	C	Unsure
23	46.229.xxx.xx	26,408	A	Yes
24	68.180.xxx.xxx	25,803	A	Yes (Yahoo)
25	216.244.xx.xxx	20,600	C	Yes (Wowrack.com)
26	194.116.xx.xxx	19,904	C	Unsure
27	17.142.xxx.xxx	18,258	A	Yes (Apple)
28	216.244.xx.xxx	13,996	C	Yes (Wowrack.com)
29	216.244.xx.xxx	13,054	C	Yes (Wowrack.com)
30	160.45.xxx.xxx	12,520	B	No (Stanford researchers)

Table 15. IP-addresses that used DBpedia-NL excluding class A (anonymous IP-addresses)

Ip-addresses that used DBpedia-NL without class A				
#	IP-address	Frequency	IP class	Bot?
1	130.251.xx.xxx	130,780	B	Unsure
2	138.201.xxx.xx	41,336	B	Unsure
3	202.180.xx.xxx	32,570	C	Unsure
4	216.244.xx.xxx	20,600	C	Yes (Wowrack.com)
5	194.116.xx.xxx	19,904	C	Unsure
6	216.244.xx.xxx	13,996	C	Yes (Wowrack.com)
7	216.244.xx.xxx	13,054	C	Yes (Wowrack.com)
8	160.45.xxx.xxx	12,520	B	No (Stanford researchers)
9	157.55.xx.xxx	8,504	B	Yes (MSN)
10	157.55.xx.xxx	7,090	B	Yes (MSN)
11	188.246.xxx.xxx	6,820	B	No
12	204.44.xx.xxx	5,840	C	Unsure
13	207.46.xx.x	5,540	C	Yes (MSN)
14	185.138.xxx.xx	4,704	B	Yes (wise-guys.nl)
15	157.55.xx.xxx	3,974	B	Yes (MSN)
16	207.46.xx.xxx	3,868	C	Yes (MSN)
17	193.206.xxx.xx	3,804	C	Unsure
18	157.55.xx.xxx	3,250	B	Yes (MSN)
19	207.46.xx.xxx	3,072	C	Yes (MSN)
20	207.46.xx.xx	2,446	C	Yes (MSN)

Fig. 5. Average usage of DBpedia-NL

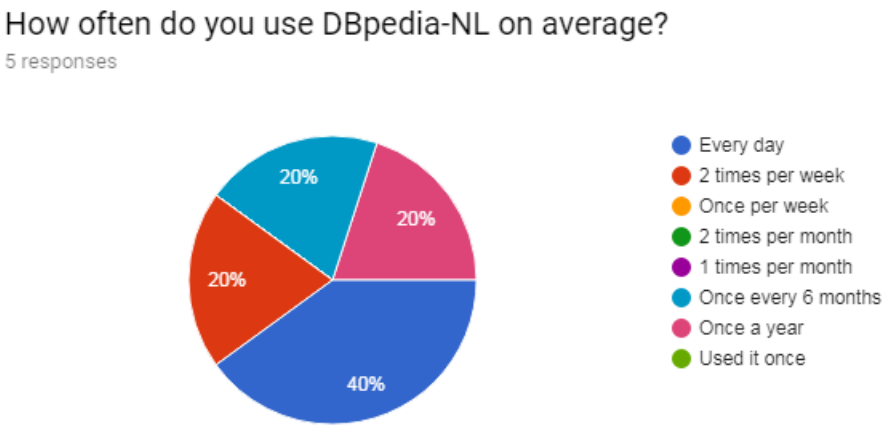


Fig. 6. How do the respondents access DBpedia-NL

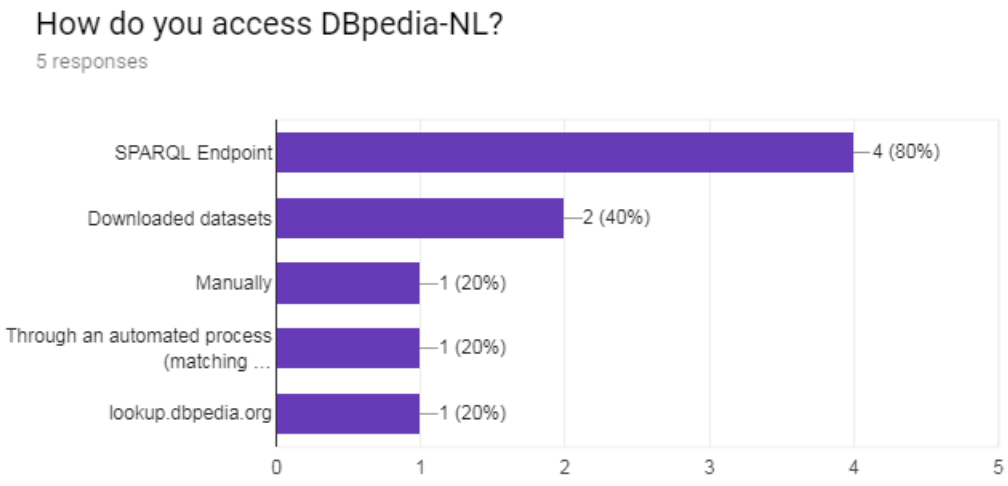


Fig. 7. What do the respondents use DBpedia-NL for

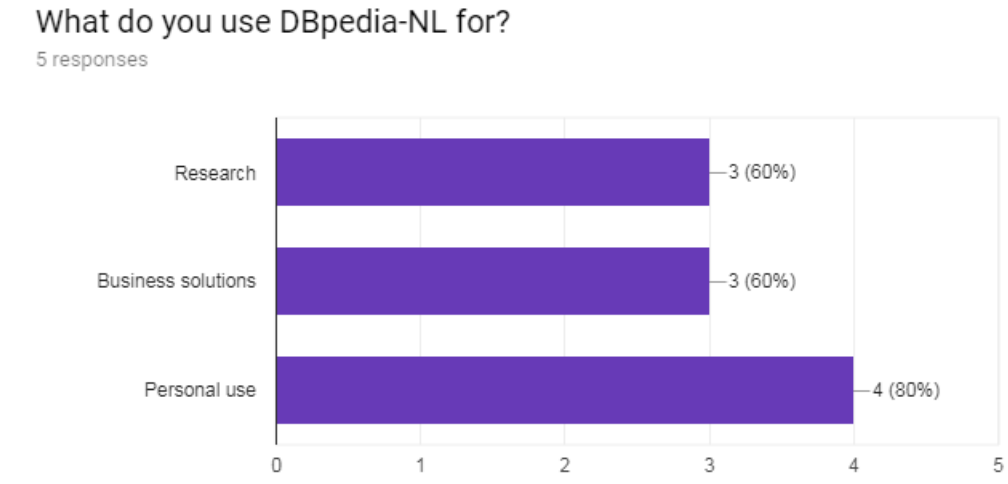


Fig. 8. Popular categories on DBpedia-NL

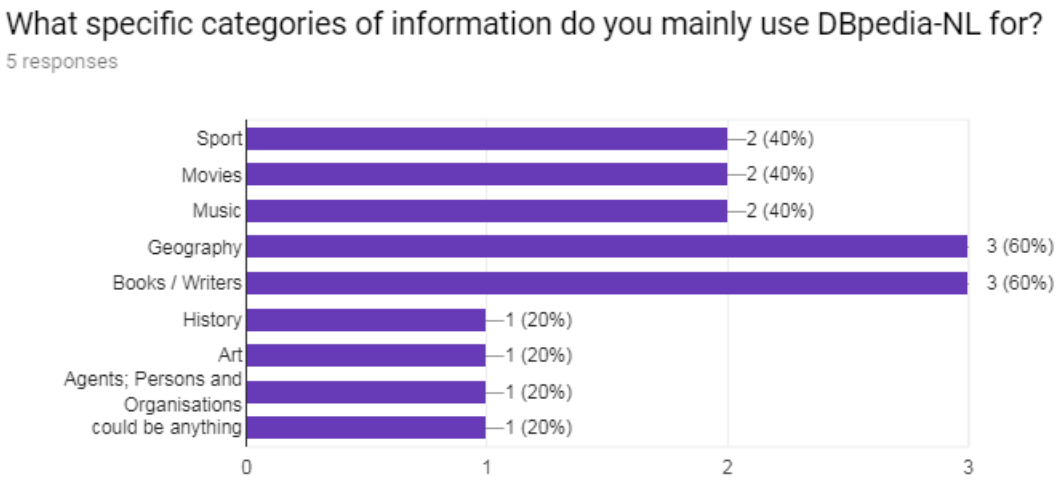
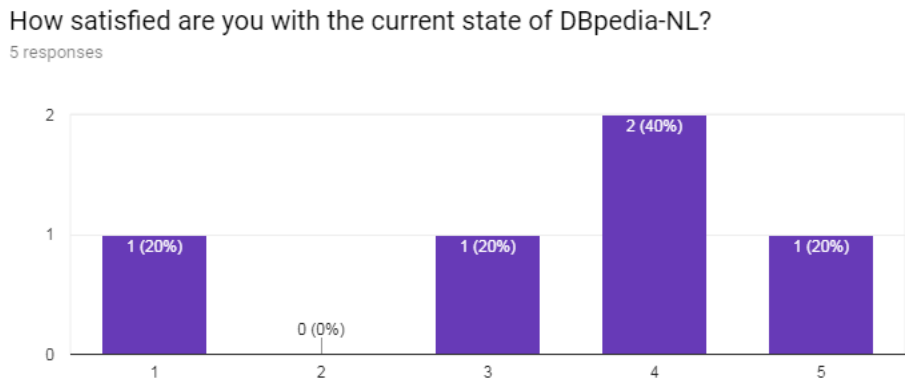


Table 16. What specific components of DBpedia-NL do you think need to be expanded on?

	What specific components of DBpedia-NL do you think need to be expanded on?
1	Better quality of data
2	More complete data and more links to other sources
3	None. I use Wikidata now instead of DBpedia
4	Links from DBpedia to (valuable) resources in datasets linking to DBpedia, i.e. backlinks
5	Did not really look much beyond the lookup API, but I mostly hope it will be integrated with wiki data first (it's so inconvenient to have two sources of data) and then I hope the Linked Data aspect will be fortified further from there, meaning: a nice consistent & simple upper ontology/schema (high quality & easy to understand by programmers/users) and subsequently a way to form communities/working groups to work on ontologies/schemas that are optimised for a certain domain (e.g. for museums or for fashion or medicine or...). A user can then decide to expand his search into different domains, by inspecting the provenance/authority of each community.

Fig. 9. How satisfied are the respondents with DBpedia-NL



REFERENCES

[1] [n. d.]. DBpedia-NL website survey publication. Retrieved June 20th, 2018 from http://nl.dbpedia.org/web/nieuws/wie-gebruikt-dbpedia
[2] [n. d.]. DBpedia WIKI. Retrieved June 20th, 2018 from http://wiki.dbpedia.org/
[3] [n. d.]. DBpedia Wiki dataset information. Retrieved June 20th, 2018 from https://wiki.dbpedia.org/services-resources/datasets/dbpedia-datasets
[4] 2004. IP-addresses explanation. https://www.bleepingcomputer.com/tutorials/ip-addresses-explained/
[5] 2018. DBpedia Usage Report. https://medium.com/virtuoso-blog/dbpedia-usage-report-as-of-2018-01-01-8cae1b81ca71
[6] DBpedia Assocation. 2018. DBpedia international datasets. https://wiki.dbpedia.org/develop/datasets
[7] DBpedia Assocation. 2018. General information. https://wiki.dbpedia.org/dbpedia-association
[8] DBpedia Assocation. 2018. Get involved. https://wiki.dbpedia.org/get-involved
[9] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. 2007. Dbpedia: A nucleus for a web of open data. *The semantic web* (2007), 722–735.
[10] P. Berkhin. 2006. A survey of clustering data mining techniques. *facilities* (2006), 25–71.
[11] P. M. Davis. 2004. InformationâĂŘseeking behavior of chemists: A transaction log analysis of referral URLs. *Journal of the Association for Information Science and Technology* 55(4) (2004), 326–332.
[12] H. Halpin. 2009. A Query-Driven Characterization of Linked Data. *LDOW* (2009).
[13] D. J. Hand. 2007. Principles of data mining. *Drug safety* 30(7) (2007), 621–622.
[14] A. Ismayilov, D. Kontokostas, S. Auer, J. Lehmann, and S. Hellmann. 2017. Wikidata through the Eyes of DBpedia. *Semantic Web, (Preprint)* (2017), 1–11.
[15] B. J. Jansen. 2006. Search log analysis: What it is, what's been done, how to do it. *Library & information science research* 28(3) (2006), 407–432.
[16] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, and C. Bizer. [n. d.]. DBpediaâĂŞa large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6(2) ([n. d.]), 169–195.
[17] L. Limam, D. Coquil, H. Kosch, and L. Brunie. 2010. Extracting user interests from search query logs: A clustering approach. *Database and Expert Systems Applications (DEXA), 2010 Workshop* (2010), 5–9.
[18] N. Sharma, A. Bajpai, and M. R. Litoriya. [n. d.]. Comparison the various clustering algorithms of weka tools. *Grouping multidimensional data* 4(7) ([n. d.]).
[19] P. Wang, M. Berry, and Y. Yang. 2003. Mining longitudinal Web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology* 54 (2003), 743–758.