Analysis of structured data on Wikipedia

Johny Moreira*, Everaldo Costa Neto and Luciano Barbosa

Centro de Informática,
Universidade Federal de Pernambuco,
Recife, Pernambuco, Brazil
Email: jms5@cin.ufpe.br
Email: ecsn@cin.ufpe.br
Email: luciano@cin.ufpe.br
*Corresponding author

Abstract: Wikipedia has been widely used for information consumption or for implementing solutions using its content. It contains primarily unstructured text about entities, but it can also contain infoboxes, which are structured attributes describing these entities. Owing to its structural nature, infoboxes have been shown useful to many applications. In this work, we perform an extensive data analysis on different aspects of Wikipedia structured data: infoboxes, templates and categories, aiming to uncover data issues and limitations, and to guide researchers in the use of these structured data. We devise a framework to process, index and query the Wikipedia data, using it to analyse different scenarios such as the popularity of infoboxes, their size distribution and usage across categories. Some of our findings are: only 54% of Wikipedia articles have infoboxes; there is a considerable amount of geographical and temporal information in infoboxes; and there is great heterogeneity of infoboxes across a same category.

Keywords: metadata; knowledge management; structured data; data analysis; Wikipedia; infoboxes; indexing strategy; categories; templates; entities.

Reference to this paper should be made as follows: Moreira, J., Costa Neto, E. and Barbosa, L. (2021) 'Analysis of structured data on Wikipedia', *Int. J. Metadata Semantics and Ontologies*, Vol. 15, No. 1, pp.71–86.

Biographical notes: Johny Moreira is a PhD Candidate in Computer Science at Universidade Federal de Pernambuco (UFPE) where he has also received his master degree. Research interests include web mining, natural language processing, machine learning and data analytics.

Everaldo Costa Neto is a Professor at Instituto Federal do Amapá (IFAP) and PhD Candidate in Computer Science at Universidade Federal de Pernambuco (UFPE) where he has also received his Master degree. His research interests include data integration, data quality, web mining and machine learning.

Luciano Barbosa is currently an Associate Professor with the Computer Science Department, Universidade Federal de Pernambuco. In addition to his experience in academia, he also held different positions in industry research labs as a Visiting Scholar Researcher at Google AI (formerly Google Research); a Research Scientist at IBM Research and AT&T Research Labs; and a PhD Summer Intern at Yahoo! Research (Europe and USA). His research interests include web mining, text mining, natural language processing, information retrieval and data analytics.

1 Introduction

Wikipedia is a collaborative, universal, and multilingual encyclopedia on the web built under the wiki principle in which any user can contribute to its content. It has become one of the best sources for creating and sharing a massive volume of human knowledge (Zhang et al., 2014).

Wikipedia is composed of a set of articles describing different entities (e.g., politicians, products, companies, artists). Each Wikipedia article contains elements such as text (organised in sections), infoboxes that describe attributes of the article's entity, images, tables and categories.

One of the reasons that makes Wikipedia content extremely valuable for applications is the structured content within infoboxes. This information has been used, for instance, to augment search engines' results, and to build knowledge bases such as YAGO (Suchanek et al., 2008; Mahdisoltani et al., 2015) and DBpedia (Lehmann et al., 2015). In addition, different studies have used infobox data on tasks such as information extraction (Lange et al., 2010; Wu and

Weld, 2017), schema clustering (Nguyen et al., 2012), multilingual schema matching (Nguyen et al., 2011; Rinser et al., 2013) and question and answering systems (Morales et al., 2016; Abbas et al., 2016).

Given the large interest on infobox data, we analyse in this work different aspects of infoboxes aiming to help the Wikipedia community to uncover some data limitations and to guide researchers and practitioners interested in performing tasks using this data.

Previous studies have already examined Wikipedia information (Wecel and Lewoniewski, 2015; Lewoniewski, 2017, 2019) by trying to assess the quality of an article through correlated features, or by analysing infoboxes across different languages. However, they usually look only at the references number, incoming and outcoming links, and the ratio of filled properties in the infoboxes. In addition, Rodriguez-Hernandez et al. (2020) proposed a tool to help Wikipedia editors to create more accurate infoboxes. Our work is complementary to them since we look not only at general aspects of infoboxes to perform properties suggestion or assess the quality of articles, but also present a process to index, query and examine infoboxes across distinct Wikipedia categories and templates.

The work of Lerner and Lomi (2018) is similar to ours in the sense of looking at Wikipedia categories. However, their investigation is restricted to the level of attention an article or category gets from users depending on which level of the Categories Hierarchy Tree they are located. In addition, beyond categories, we study Wikipedia templates, used to suggest attributes for infoboxes's creation, to verify how these templates are used to define infoboxes, and how these structures are organised across the categories hierarchy tree.

The process here reported include these different structures. The majority of our findings were not previously reported, and with the assistance of the proposed indexing and querying processes, they can be easily updated to reflect new versions of Wikipedia dump and DBpedia data sets.

Categories and templates are important sources of semantic knowledge. The hierarchy, structure, schema and even the names of categories and templates. For example, categories can assist taxonomy derivation (Kotlerman et al., 2011; Vivaldi and Rodrguez, 2010), and topic discovery through matching and clustering (Titze et al., 2014; Lalithsena et al., 2017) tasks, while template information can benefit schema discovery (Wu and Weld, 2017).

In this study, we analyse different aspects of Wikipedia data such as: the size of the infoboxes and popularity of the attributes and templates; the distribution of spatial and temporal attributes, and templates in different categories; the usage of attributes of the templates in infoboxes; and the heterogeneity of infoboxes in the same category.

Many studies can benefit from our work. For instance, knowing the rich coverage of geographical and temporal information can help investigations in the spatial-temporal area. And learning issues such as: infoboxes that use too generic templates contain a large number of unused properties, and different templates in the same category share similar attributes, can serve as input to the Wikipedia editors to produce better content and standards.

To perform this data analysis, we implemented a pipeline for pre-processing, building and searching Wikipedia's structured data, built from Wikipedia and DBpedia data sets. We use three DBpedia data sets to recover the mappings of articles to categories, categories to subcategories and attributes (parameters) of suggested templates. To recover the data related to infobox instances we parse the Wikipedia dump. We perform a pre-processing step of the DBpedia data sets triplets to simplify the indexing and querying of information. The infobox instances and template mappings retrieved from Wikipedia dump are also indexed. After indexing all the mappings between these structured data, we are able to retrieve general information as well as information related to specific domains to perform the analysis. To retrieve data related to a given domain of information we propose a querying scheme that retrieves data over categories and subcategories. The resulting indexes and search engine are available to download.

The remainder of the paper is organised as follows: in Section 2, we list and compare previous studies, showing highlights of their findings. We describe in Section 3 our strategy to Wikipedia structured data and the framework to index and query it. Using this framework, in Section 4, we first provide some general statistics about the data. Thus, in Section 5, we analyse some aspects of infoboxes across different categories. In Section 6, we present a summary of findings, discussing scenarios where our study can contribute. We conclude in Section 7 with a summary of our findings.

2 Related work

Related to our study, Wecel and Lewoniewski (2015) evaluated the hypothesis that the overall quality of Wikipedia articles allows the derivation of quality measures for attributes in infoboxes and vice-versa, focusing specially on infoboxes attributes. For that, they analyse features and models to evaluate the quality of articles. The result of the analysis are positively correlated features that are used as input for automatic quality classification of articles. Some examples of these features are: attributes' length, the number of references, images, editions, and incoming links. As prediction targets, they use the Wikipedia community quality classes:[1]: "Featured Article" (*FA*), "Good Article" (*GA*), *A*, *B*, *C*, *Start* and *Stub* to articles. Classes *A*, *B*, and *C* are used to classify intermediate articles, while *Start* points out to articles under developing and incomplete; *Stub* indicates articles missing relevant information. The highest classes (*FA* and *GA*) are assigned only after discussion and agreement inside the community, whereas the others can be assigned by regular users. They performed their study over six different language versions of Wikipedia. Owing the lack of the complete grading scale in some languages, they split the articles for building the models in two major classes: 1 – complete (classes FA and GA); and 2 – developing (encompassing all other classes). Their results show that the quality of infobox attributes has influence on the overall quality of the article, and that each language presents features with different significance. This work is related to ours in the sense that they leverage structured information for quality assessment of Wikipedia infoboxes.

Following the work of Wecel and Lewoniewski (2015); Lewoniewski (2017, 2019) also analysed different aspects of

articles and infoboxes across different languages. Lewoniewski (2017) focused on the study of two quality dimensions: completeness and reliability of the infoboxes. Completeness is the ratio of the number of attributes in an infobox over the total number of properties of the infobox's template. The reliability is composed of three metrics: number of outlinks, unique references, and the ratio of filled properties with reference over the total number of references. The analysis is performed over seven different language versions of Wikipedia, and each language is analysed in five different selected topics: *Album*, *Companies*, *Films*, *Universities* and *Video Games*. Regarding the completeness quality measure, there is a great variation across languages. For instance, the average weighted completeness of infoboxes for *Album* in English is 0.452 while in French is 0.223. Regarding reliability, the study discovers that in certain domains some languages use few outlinks in the infobox. As an example, only 0.8% of infoboxes in the French *Films* domain have oulinks. Lewoniewski (2019) extended the previous studies by outlining other dimensions for articles and infoboxes to measure and evaluate the quality of Wikipedia. Some examples of these dimensions are: *relevance*, measured by the number of unique authors of the infobox (or if it was built by bots, anonymous users or administrators); and *timeliness*, which is related to the number of recent changes in the infobox. Their work only describes these dimensions, but no evaluation or data analysis is performed.

Another work that assesses the quality of structured data on Wikipedia, proposed by Reznik and Shatalov (2016), investigates biographical data from Wikipedia pages using information present on infobox structures and the number of links to a page as a measure of significance. The study showed people's interest in historical figures over time, confirming that Wikipedia content has a good potential for studies related to temporal information.

Lerner and Lomi (2018) performed a data analysis using statistical methods over Wikipedia data. The study investigates how the position of a category in the categories hierarchy can attract more or less attention from Wikipedia editors and how its position can affect the quality evaluations of articles classified under this category. One of the findings is that articles under coarse-grained categories (located high in the hierarchy) are more likely to be popular, but less likely to be evaluated as of high quality.

More recently, Guda et al. (2020) proposed NwQM, a model similar to Wecel and Lewoniewski (2015) for automatic assessment of Wikipedia articles quality. However, they apply as features only the article text, images and some metadata. For that, NwQM builds a deep learning model for document classification assigning the classes *FA*, *GA*, *B*, *C*, *Start* and *Stub*. This work makes use of some meta contents such as infobox data and categories as special tokens to use it as additional features for classification. Although this work is build upon the articles quality problem, it does not directly relate to ours. NwQM uses a document classification approach to automatically assign quality classes while ours performs a data analysis of Wikipedia structured data in order to give insights on its structure and data distribution. This observation also applies to the other studies described in this section performing automatic quality classification for articles.

The work of Rodriguez-Hernandez et al. (2020) presents WikInfoboxer, a tool that considers infobox templates to perform suggestions for Wikipedia editors to create richer and more accurate infoboxes. For this, it aggregates and ranks similar templates and properties to suggest to editors. Whenever possible, the tool also suggests values for the properties and links the new created infoboxes to the respective entity. Although, WikInfoboxer also uses DBpedia data and deals with infobox creation considering infobox templates, our works differ in the sense that we present a process to access and evaluate infobox information accounting for the categories' hierarchy.

3 Data processing

To perform our study on Wikipedia infoboxes, we used a combination of DBpedia (Lehmann et al., 2015), a knowledge base created from Wikipedia, and Wikipedia data sets. In this section, we provide details about the data and our strategy to process it for the analyses.

3.1 Data description

DBpedia is a knowledge base represented by Resource Description Framework (RDF) triples, created from Wikipedia structured information. It provides information about entities, categories of the entities, subcategories, hyperlinks between entities etc. For this study, we used the English dump of October of 2016[2], which comprises 6.6 M entities and a total of 1.7 billion of RDF triples. It also contains the NLP Interchange Format (NIF) annotation (Hellmann et al., 2013), which enables interoperability between NLP tools, language resources and annotations. The use of NIF by DBpedia makes it possible to include in the data sets the whole wiki text, its basic structure (sections, titles, paragraphs, etc.) and the links.

Each data set is composed of a series of turtle files (.ttl) containing triples in the format *subject-predicate-object* $(< s >< p >< o >)$. For this particular study, we used the English version of the following DBPedia data sets:

article-categories data set holds relations between articles and categories (or subcategories) in triples (23,990,512 in total). The structure of its triples is organised as follows: the *subject* corresponds to an article Uniform Resource Identifier (URI), the *predicate* is a property and the *object* corresponds to the category URI. Listing 1(a) shows an example of a triple in this data set.

- *skos-categories* data set represents relations between categories and subcategories, consisting of 6,083,029 triples. This data set uses the property *broader* from the Simple Knowledge Organisation System (SKOS)[3] namespace document, which indicates a subcategory-category relation in the triple. The *subject* corresponds to the subcategory, the *predicate* holds the *broader* annotation and the *object* corresponds to a category. Listing 1(b) shows an example of a triple in this data set.

The Wikipedia Community also provides information about infobox templates[4], which should be used in the creation or

editing of infoboxes. As described by the community itself[5], the infobox templates contain suggested attributes in a particular topic to provide standardised information across related articles. When creating or editing an article, the user should inform its template at the infobox mapping header. To perform the analysis on infobox templates, we used the *template-parameters* data set available on DBpedia. It contains triples that consist of the template URI (subject), URI for the ontology property *templateUsesParameter* (predicate) and the template parameter name (object). Listing 1(c) shows an example of a triple in this data set. This file comprises a total of 776.554 triples.

Listing 1 Tuple examples for each dataset along with its respective processing output

(a) article_categories_en.ttl

```
(
    <http://dbpedia.org/resource/Tow_Law>
    <http://purl.org/dc/terms/subject>
    <http://dbpedia.org/resource/Category:Wind_farms_in_England> .
)
Indexing Terms:
Subject = Tow_Law
Predicate = subject
Object = Wind_farms_in_England
```

(b) skos_categories_en.ttl

```
(
    <http://dbpedia.org/resource/Category:Wind_farms_in_England>
    <http://www.w3.org/2004/02/skos/core#broader>
    <http://dbpedia.org/resource/Category:Wind_farms_in_the_United_Kingdom>
)
Indexing Terms:
Subject = Wind_farms_in_England
Predicate = none
Object = Wind_farms_in_the_United_Kingdom
```

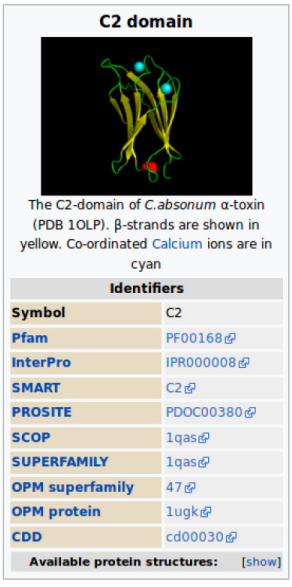
(c) template_parameters_en.ttl

```
(
    <http://en.dbpedia.org/resource/Template:Infobox_UK_place>
    <http://en.dbpedia.org/property/templateUsesParameter>
    "latitude" .
)
Indexing Terms:
Subject = Infobox_UK_place
Predicate = templateUsesParameter
Object = latitude
```

Figure 1 Infobox types

(a) infobox

(b) pfam_box



(c) taxobox

(d) chembox

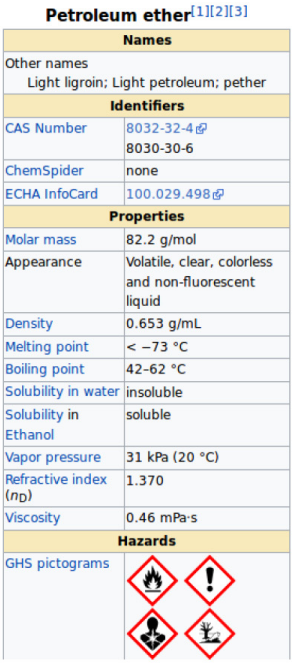


Figure 1 presents examples of infoboxes using different templates: *infobox*, *pfam_box*, *taxobox* and *chembox*. Although their structures look similar, their mappings change according to the template type. Some of the templates are related to specific domains (e.g., *chembox* is associated to the chemicals and *pfam_box* to protein family or domain) whereas the *infobox* template, for instance, is used across topics.

To understand the popularity of the templates on our data set, we calculated the proportion of the types of templates using the regular expressions, present in Table 1 to identify them. Templates starting with *"Infobox"* are the most common ones: 88.4% of the total. This might indicate the result of an effort by the Wikipedia community to standardise the templates towards the pattern. Based on this, in this work we have considered only the infobox template mapped with prefix *"Infobox_"*.

Table 1 Count occurrences of infoboxes types

Infobox pattern in regex	*Proportion*
{{\s?(I|i)nfobox.*\n(|.*\n)*}}	88.49×10^{-2}
{{\s?(T|t)axobox.*\n(|.*\n)*}}	9.48×10^{-2}
{{\s?(S|s)peciesbox.*\n(|.*\n)*}}	52.72×10^{-4}
{{\s?(G|g)eobox.*\n(|.*\n)*}}	52.23×10^{-4}
{{\s?(C|c)hembox.*\n(|.*\n)*}}	37.71×10^{-4}
{{\s?(A|a)utomatic(_|\s)taxobox.*\n(|.*\n)*}}	23.27×10^{-4}
{{\s?(D|d)rugbox.*\n(|.*\n)*}}	22.47×10^{-4}
{{\s?(E|e)nzyme.*\n(|.*\n)*}}	12.04×10^{-4}
{{\s?(P|p)fam(_|\s)?box.*\n(|.*\n)*}}	02.06×10^{-4}
{{\s?(P|p)rotein.*\n(|.*\n)*}}	37.89×10^{-6}

In this study, we are also interested in the structural information (attribute-value pairs) of the infoboxes. To obtain this, we initially tried to use DBpedia data sets. However, these data sets contain noisy information (e.g., properties not present in the article's infobox such as image frames) and lack coverage with respect to infoboxes' attributes. For this reason, we opted to use the English Wikipedia dump (dating from October 2016, same as the DBpedia dump) in order to extract infoboxes, infoboxes template mappings, infobox properties and its respective values. The Wikipedia dump consists of an XML file containing all pages from the given revision. Each Wikipedia page is written in *Wikicode*, a markup language for formatting Wiki pages from the *MediaWiki Foundation*. For the extraction of infobox data we used a parser for wikicode[6].

3.2 Indexing strategy

To efficiently query the data sets and be able to generate statistics for this study, we used Apache Lucene7 to index the DBPedia and Wikipedia data sets. We index the DBPedia data sets iterating over each data set triple, considering it as a document and each one of its elements as a term. The predicate statement, however, was not used for indexing since it does not contain any relevant information for our study. During indexing of these triples, we decided to keep only significant information for better readability. For that, we tokenise each triple element by removing part of the URIs and special characters like: quotes, ">" and "<", selecting for indexing only the last token from URIs in triple statements, as presented on Listing 1.

There is no need to remove underscores "_" from indexed terms since the applied indexing analyser treats the entire stream as a single token. This approach allows searching by exact match and ensures the consistency of the data analysis. Also, the underscore "_" is present in all analysed Wikipedia structures (categories, templates, templates parameters and infobox properties), and the aim of our work is to analyse the organisation of these raw data. Hence, we decide for not removing it from the indexing terms since it makes intuitive the arrangement of the original URIs when necessary.

For the Wikipedia dump, we applied the following strategy: we iterate over each tuple from an existing infobox's article, indexing it as a triple: the article name as subject, the property name as predicate and the property value as object. We also index the infobox template mapping as a tuple: subject (article name) and object (infobox template mapping).

A *Keyword Analyser* is applied while indexing the terms to keep the index and query results trustworthy, without missing any relevant information. During indexing and querying, we do not apply tokenisation, stemming, stop-word or any other type of filter to ensure the indexed triplets are rightfully queried.

Table 2 presents an overview of the data after processing. There are 5.166.304 articles in 1.079.614 categories. Around half of the articles contains infoboxes (2.7 million), and there are 56.819 unique properties (or attributes) in those infoboxes. Regarding mapped templates to infoboxes with the *"Infobox"* prefix, there are only 3.448 different templates.

The indexes built for this work are available to download[8] and can be easily consumed via Apache Lucene API (Version 6.6.2). A basic query engine for indexes exploration is available as well[9].

Table 2 Overview of our data set

	Total
Articles	5.166.304
Infoboxes	2.785.031
Categories	1.263.435
Categories with articles	1.079.614
Properties in infobox instances	56.819
Mapped infobox templates	3.448

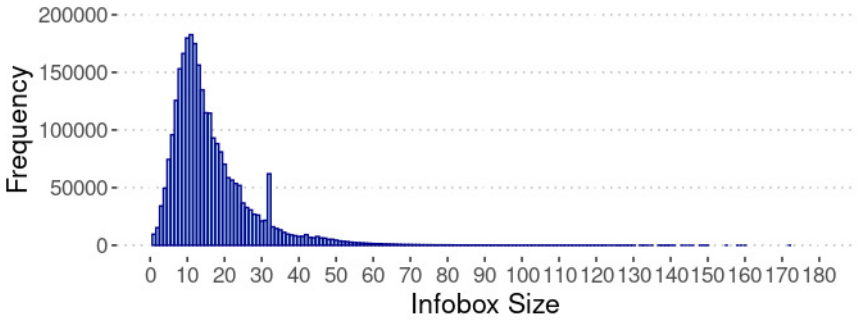
4 General statistics

In this section, we provide statistics about the structure and content of Wikipedia infoboxes and infobox-template mappings on our data.

4.1 Size of the infoboxes

Figure 2 shows the distribution of the size of infoboxes on Wikipedia. The infobox size is measured by counting the number of distinct properties. Although the numbers show a long tail distribution of infoboxes size there is still a reasonable number of properties per infobox: its median size of is 13. The spike in Figure 2 around size 32 comes mostly from entities instantiating the "*Infobox Settlement*" template (52.545 instances). This template is the most used one in our data set and one of the largest in terms of number of suggested properties (424), as we further detail in the next sections.

Figure 2 Distribution of infoboxes size on Wikipedia



The largest infobox, with 172 properties, corresponds to the *Al_Jazeera_English*[10] entity, a Qatari paid television news channel. This entity instantiates the infobox template "*Infobox_television_channel*"[11], which also presents a large number of suggested properties (328). Mostly, the properties are related to variations of satellite, cable, online, IPTV services, and channel number information. Although these properties represent a same concept with slightly variations, they are suggested as individual attributes, e.g.: "*sat_serv_1*" going from 1 to 30, "*sat_chan_1*" also going from 1 to 30, "*cable_serv_1*" going from 1 to 20, among others. These properties could be replaced, for instance, by multivariate ones which would define a more concise and homogeneous template. The second biggest infobox, with 160 properties, is related to *Jefferson_Louis*, a British soccer player. Its infobox instantiates the "*Infobox_football_biography*" template[12] which is the largest template in our data set with 426 suggested properties. This template also specifies multivariate properties as individual properties, e.g. Youth career, College career and Senior Career, which are presented in Figure 3. We can conclude from these observations that the number of attributes of an infobox template has a great influence in the infobox size of articles that use it, even though the template's attributes are not required.

Figure 3 Infobox template for soccer player's or manager's

{{{name}}}

[[File:{{{image}}}|{{{image_size}}}|alt={{{alt}}}|{{{alt}}}]]
{{{caption}}}

Personal information	
Full name	{{{fullname}}}
Birth name	{{{birth_name}}}
Date of birth	{{{birth_date}}}
Place of birth	{{{birth_place}}}
Date of death	{{{death_date}}}
Place of death	{{{death_place}}}
Height	{{{height}}}
Playing position	{{{position}}}
Club information	
Current team	{{{currentclub}}}
Number	{{{clubnumber}}}
Youth career	
{{{youthyears1}}}	{{{youthclubs1}}}
{{{youthyears2}}}	{{{youthclubs2}}}
{{{youthyears3}}}	{{{youthclubs3}}}
{{{youthyears4}}}	{{{youthclubs4}}}
{{{youthyears5}}}	{{{youthclubs5}}}
{{{youthyears6}}}	{{{youthclubs6}}}
{{{youthyears7}}}	{{{youthclubs7}}}
{{{youthyears8}}}	{{{youthclubs8}}}
{{{youthyears9}}}	{{{youthclubs9}}}
{{{youthyears10}}}	{{{youthclubs10}}}

College career			
Years	**Team**	**Apps**	**(Gls)**
{{{collegeyears1}}}	{{{college1}}}	{{{collegecaps1}}}	({{{collegegoals1}}})
{{{collegeyears2}}}	{{{college2}}}	{{{collegecaps2}}}	({{{collegegoals2}}})
{{{collegeyears3}}}	{{{college3}}}	{{{collegecaps3}}}	({{{collegegoals3}}})

Senior career*			
Years	**Team**	**Apps**	**(Gls)**
{{{years1}}}	{{{clubs1}}}	{{{caps1}}}	({{{goals1}}})
{{{years2}}}	{{{clubs2}}}	{{{caps2}}}	({{{goals2}}})
{{{years3}}}	{{{clubs3}}}	{{{caps3}}}	({{{goals3}}})
{{{years4}}}	{{{clubs4}}}	{{{caps4}}}	({{{goals4}}})

4.2 Most popular properties

Figure 4 presents a word cloud with the most popular properties in the infoboxes. As can be seen, some of them are from generic entities. For instance, *nxame*, *birth_place* and *birth_date* are attributes of the Person entity. Attributes related to geographical and temporal information are also common: e.g., *location*, *longitude*, *latitude*, *coordinates*, *timezone*, *country*, *birth_place*, *birth_date*, *death_date* and *years*. In addition, there is a great occurrence of properties representing similar semantic information with generic names: *subdivision_type*, *subdivision_type1* and *subdivision_type2*.

4.3 Most popular infobox templates

Figure 5 presents the number of articles that use the top-20 templates for the "*Infobox*" prefix. Some of them refer to high-level entities which represent broader concepts such as *settlement*, *person*, *company*, *officeholder* and *station*. The top template *Infobox_settlement*[13] is recommended by Wikipedia to be used for "any subdivision below the level of a country". Other templates on the top-20 are also related to organisations or location: company, National Register of Historic Places (NRHP), French commune, station, school and UK place. There are also other groups on the top-20 related to people (person, football biography, musical artist, officeholder, sportsperson, military person, scientist and writer), and associated to entertainment and arts (film, album, single, television and musical artist).

Looking at the attributes of the templates in those groups, we found that same properties are used by different templates. For instance, in all templates related to people, attributes such as *birth place* and *death place* are present. A similar trend was observed in the other groups. This means that some effort is still needed to standardise templates on Wikipedia to avoid this redundancy.

Figure 4 Top property names used on Wikipedia

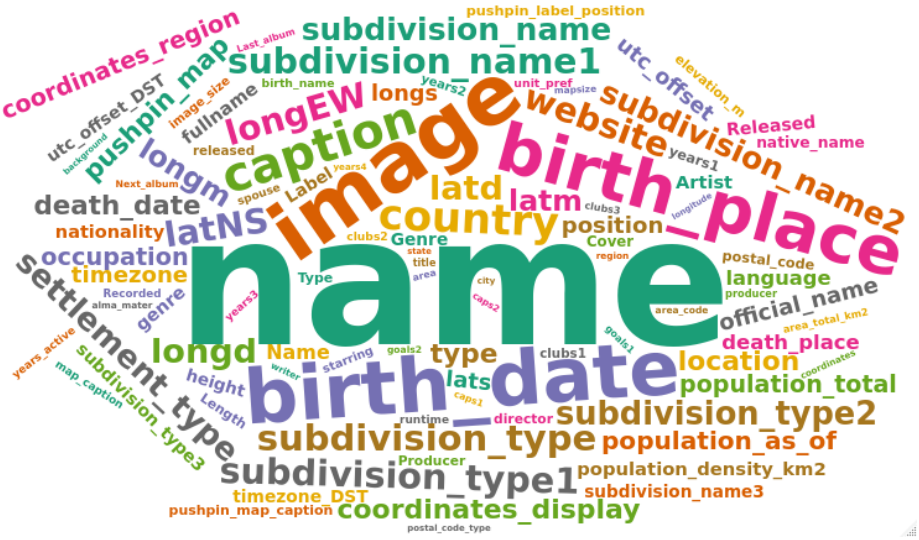
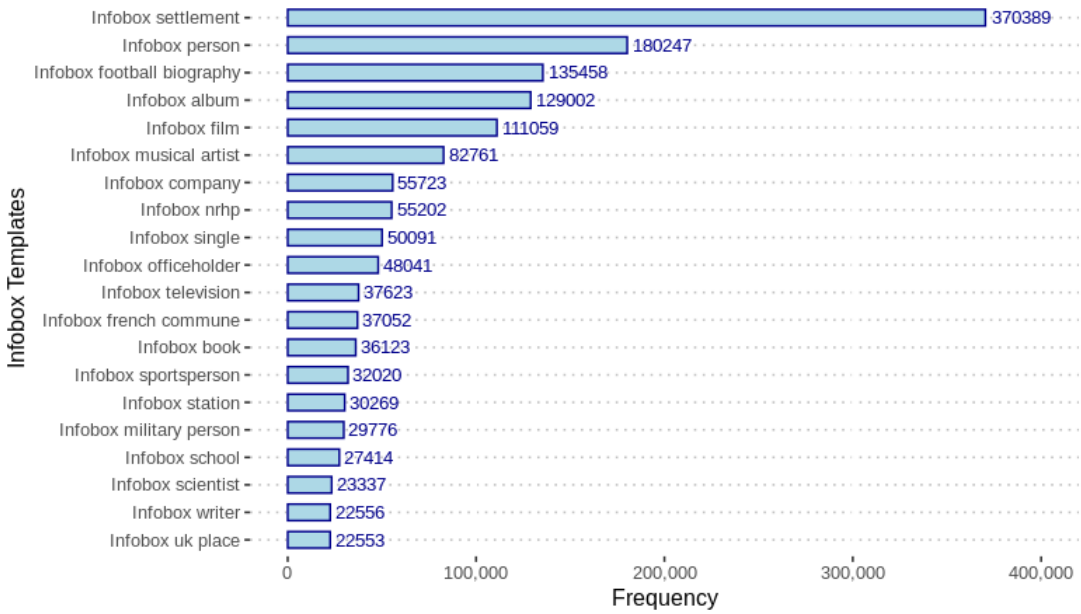


Figure 5 Top 20 retrieved infobox templates by frequency usage



5 Category-based statistics

In this section, our goal is to study how the structured information on Wikipedia is organised in different categories. Every Wikipedia article is assigned to one or more Wikipedia categories for purpose of organisation. A Wikipedia category aggregates articles related to a same topic or information domain. Additionally, a category can be associated to another building a *subcategory* relationship. This categories hierarchy is known as *Wikipedia Category Graph* (Heist and Paulheim, 2019).

Since our Wikipedia dump has more than 1 million categories, to perform this analysis, we selected 14 diverse categories. To choose them, we ranked all Wikipedia categories based on the number of articles (subcategories were not considered). Among the top-10 categories on infobox frequency there are many categories under the same topic, for instance: *Association football midfielders*, *Association football defenders*, *English Football League players*, *Association football forwards*, and *English footballers*. Hence, we manually selected random categories which correspond to different domains but also contains a considerable number of infoboxes. Table 3 shows the categories used in this study.

To collect the entities for each category, we implemented the following strategy (presented in Figure 4). Given a query that represents a Wikipedia category C, the system searches on the Categories index for articles belonging to C. We observe that the amount of entity articles in some categories can be small, since the chosen category C most times can be a hub for other categories or represent a general subject. Hence, we expand our search in order to consider subcategories of C and entity articles directly associated to C. Since the number of those entity articles might be small, the system collects entity articles of subcategories of C by navigating deeper in the category structure. For each level in the hierarchy category, the algorithm merges the infobox attributes of their articles, creating some kind of category schema, and then measures the Jaccard similarity (Jaccard, 1901) between the current category schema (.i.e., the set of all its attributes) and the schema of each child category.

Figure 6 Querying and retrieving scheme

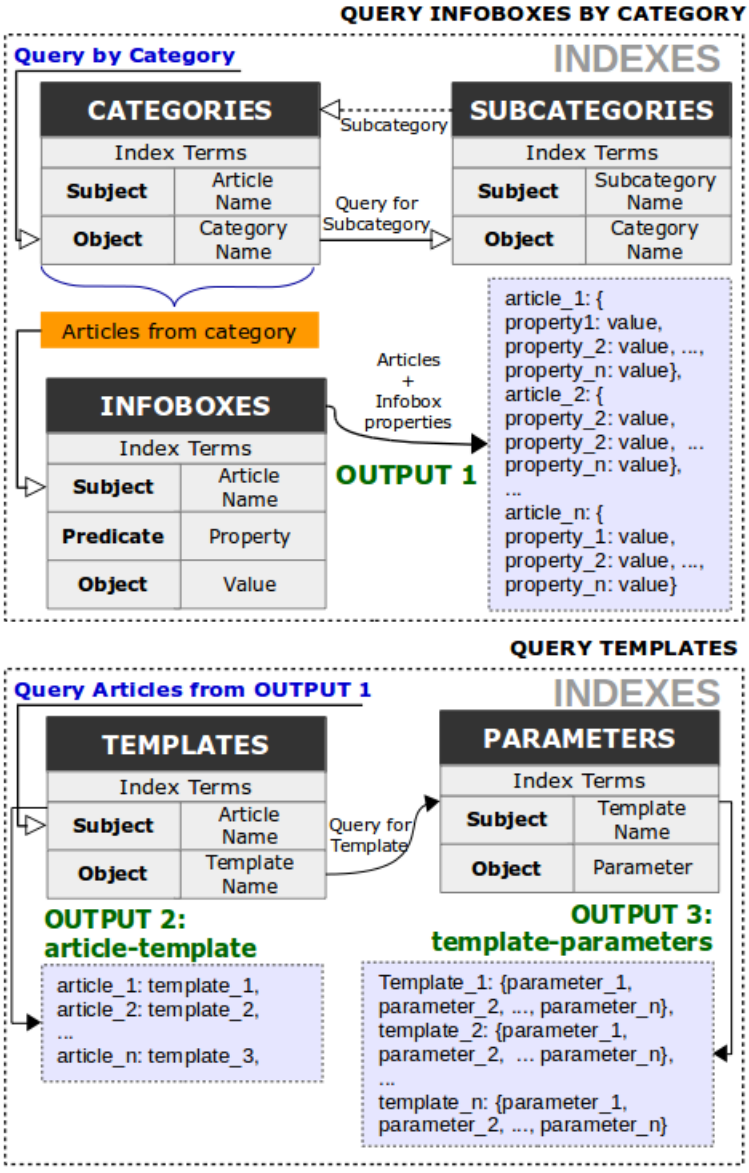


Table 3 Extracted categories, number of articles found, infobox instances, subcategory extraction nodes and a brief description of category subject

Category	*Art.*	*Inf.*	*Nodes*	*Description*
Alpine three-thousanders	944	934	2	Mountains between 3000 metres (9842 ft) and 3999 metres (13,122 ft) above sea level in the Alps, in Austria, Switzerland, Italy and France
Oil pipelines by country	105	72	123	A set of long-distance pipes for oil transportation organised by country.
Oil fields by country	558	454	146	Land areas with an abundance of oil wells extracting petroleum (crude oil) from below ground. Organised by country.
Protein families	742	379	1	Collects articles describing sets of related proteins called a family.
Skyscrapers between 100 and 149 meters	1028	905	1	Tall buildings between 100 and 149 metres.
Numbered minor planets	2959	2910	1	Articles on numbered Minor Planets (MPs)
Towns in Turkey	750	693	8	Towns in Turkey
Martial arts by type	7572	4910	359	Martial arts organised by types.
Formula One cars	680	311	140	Cars intended to be used in competition at Formula One racing events.
HarperCollins books	1832	1597	15	Books published by HarperCollins and its imprints – a subsidiary of News Corp, based in New York City.
French films	7258	6904	638	All French films.
Oil companies of the USA	464	288	20	Companies from United States Oil Industry.
Operas	3636	1008	734	Operas, subcategorised by composer, genre, original language, year and acts.
IOS games	1799	1730	8	Video games available for the iOS operating system.

The search process stops when this value is below a given threshold θ (for this study we define $\theta = 0.15$). The threshold was selected through an empirical study where we concluded that defining a larger threshold could restrict the navigation across subcategories. This is because for each category level a global schema is defined, hence there can be a large number of distinct attributes which consequently decreases the similarity between mother and child categories. On the other hand, if a threshold is not applied there could be cases in which the categories hierarchy falls into a cycle. Assigning a small threshold is just a guarantee that the search will stop while retrieving a considerable number of infobox instances.

Table 3 shows for each selected category, the number of articles, infoboxes and nodes in the hierarchy, and a brief description of each category domain collected by this process.

Next, the pipeline extracts the attribute-value pairs from the infobox's articles, and retrieves template mappings and schemata. The templates used by the articles extracted on the first step are retrieved from the *Templates* index and the schemata of those templates are then retrieved from the *Parameters* index, as presented in Figure 6.

5.1 Size of infoboxes per category

Figure 7 shows the distribution of the size of the infoboxes for the 14 categories. The median size of the infoboxes per category varies from 7 for Protein families, to 33 for Towns in Turkey. Infoboxes in Towns in Turkey use the "*Infobox Settlement*" template, which has a total of 424 attributes, whereas the Protein families category instantiates in most of the cases the *Infobox protein family* template, which is the smallest one with 21 suggested attributes among the 14 categories. This confirms a previous observation regarding the strong relation between the number of attributes of a template and the size of infoboxes that use it. Figure 7 also shows that there is not much variation in terms of the size of infoboxes within a category. The largest variation is "*Skyscrapers between 100 metres and 149 metres*" with interquartile range of 8.

In Figure 7, it is also possible to note that all categories contain outliers in terms of infobox size. Examples of the biggest infobox for each selected categories are depicted in Table 4. Among them, the category *Martial arts by type* presents the biggest infobox with 84 properties related to *Solomon Haumono*, a former professional boxer and rugby player.

Figure 7 Distribution of infobox size by category

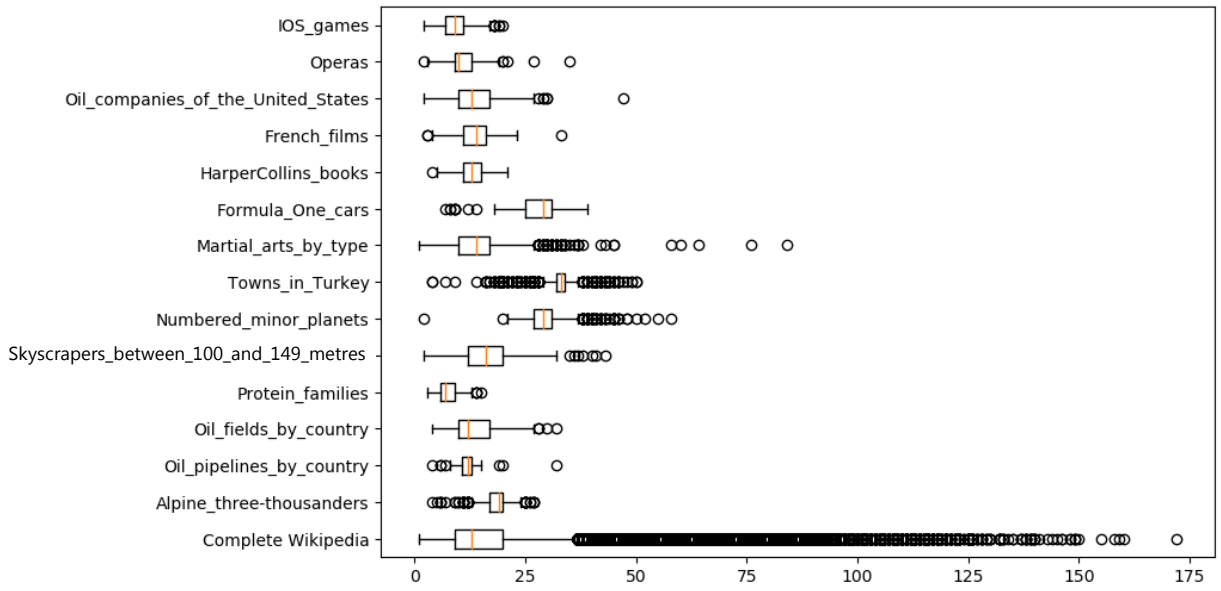


Table 4 Biggest infobox schema for each selected category

Category	*Article*	*Size*
Alpine_three-thousanders	Breithorn_(Lauterbrunnen)	27
Oil_pipelines_by_country	Trans-Alaska_Pipeline_Authorization_Act	32
Oil_fields_by_country	Boca_de_Jaruco	32
Protein_families	Fructose_1,6-bisphosphatase	15
Skyscrapers_between_100_and_149_metres	1501_Broadway	43
Numbered_minor_planets	Ceres_(dwarf_planet)	58
Towns_in_Turkey	ElmalÄ±	50
Martial_arts_by_type	Solomon_Haumono	84
Formula_One_cars	Honda_RA106	39
HarperCollins_books	The_Final_Unfinished_Voyage_of_Jack_Aubrey	21
French_films	Transporter:_The_Series	33
Oil_companies_of_the_United_States	Sunray,_Texas	47
Operas	471_Papagena	35
IOS_games	Splatterhouse	20

5.2 Spatial and temporal properties

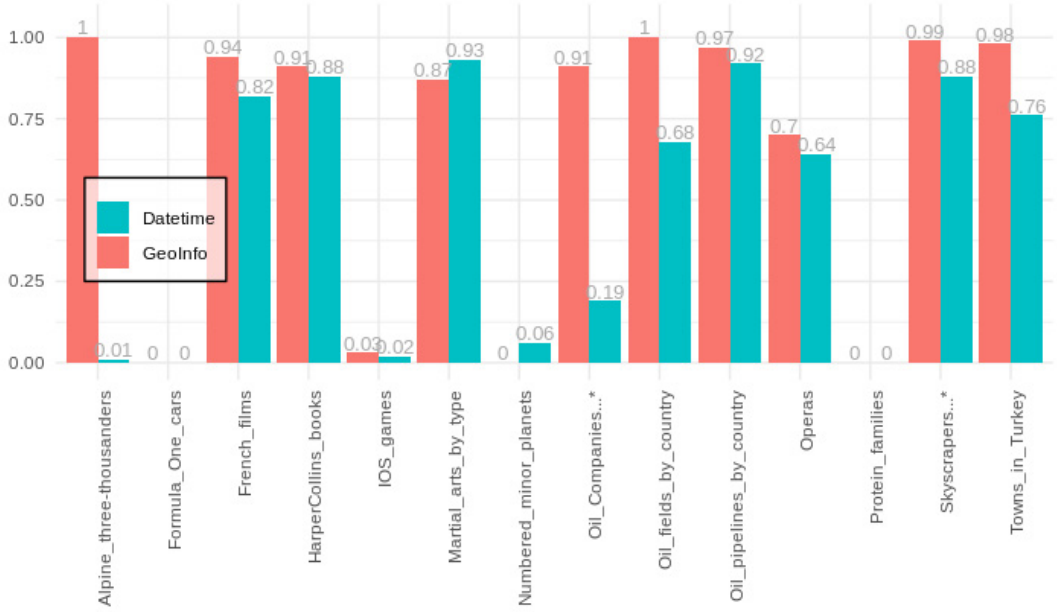
To gather spatial and temporal attribute distribution across infoboxes, we built type detectors using the rules described in Table 5. To study the spatial types, we have defined four categories: latitude, longitude, coordinates and location. Although latitude, longitude and coordinates represent the same semantic information, it is common to find properties related to each one of these classes in the same category. Location is mostly related to addresses and entity's region or origin. Regarding the temporal information, we classified it in three different types: time, time interval and date.

Overall, there is a total number of 2,105,172 infoboxes with geo information and 1,584,340 infoboxes with temporal information. Figure 8 shows the distribution of geographic and temporal information across the categories.

Table 5 Geo-time regex rules

Spatial rules	
Attribute	*Regex rules*
latitude	lat.*
longitude	lon.* | long.*
coordinates	.*coordinate.* | .*coord.*
location	*location.* | .*region.* | .*place.* (.*\ W)city.* | .*address.* .*country.* | .*residence.* .*origin.* | .*state.*
Temporal rules	
Attribute	*Regex rules*
date	.*date.* | .*year.* | .*day.* | .*month.*
time	.*time.* | .*timezone.* | .*timestamp.*
time interval	.*start.* | .*end.* | .*stop.* | .*duration.*

Figure 8 Proportion of infoboxes with spatial and temporal information



Regarding geographic information, most of the articles in the categories *Alpine three-thousanders, Oil fields by country, Oil pipelines by country, Skyscrapers between 100 metres and 149 metres*, and *Towns in Turkey* have spatial attributes. This is expected since those categories are highly related to geographic entities. Interestingly, some non-obvious categories contain a high proportion of articles with spatial information as well: *French films, HarperCollins books, and Martial arts by type*. This occurs because many infobox articles of these categories refer to infobox templates that have attributes related to location. For instance, the attribute country appears in the templates: *Infobox_film, Infobox_book for French Films* and *HarperCollins books*. Also, the attributes birth and death place in the template *Infobox_boxer* in the category *Martial arts by type*. On the other hand, it is reasonable the absence of spatial information on *Protein families* and *numbered minor planets* categories since its infoboxes are composed mainly of scientific descriptions. Geographic information is not available for *Formula One Cars* as well since the main subject of this category are racing cars' characteristics.

In terms of temporal attributes, Figure 8 also shows the proportion of infoboxes across categories. Looking at the categories with the highest proportion of temporal information, we can find a high influence of templates in the infoboxes' composition. For instance, the most used template in the *French Films* category contains the attributes running time and release date; for *Towns in Turkey* the property timezone; *Oil fields* and *Oil pipelines by country* contain the attributes time interval attributes as start, start_development, start_production properties related to date development and production started. On the other hand, the categories *Formula One Cars* and *Protein Families* do not present any temporal information. Inspecting the *Infobox racing car* template, the most used one in *Formula One Cars* category, we found time-related properties such as First_win, Last_win, Last_event, which contain date-time information. These properties, however, are composed of tokens not present in the geo-time regex rules in Table 5.

5.3 Usage of templates

Table 6 presents numbers regarding the usage of templates across categories. In almost all categories a single template is used for most of the articles: the most used template in each category covers about 90% of the articles. The exceptions are: *Martial Arts by Type* (62%), *Oil companies of the United States* (64%), *Oil fields by country* (58%), and *Operas* (37%) which also present the greatest number of templates: 96, 29, 21 and 22, respectively. This might show that these latter categories are more heterogeneous, but might also indicate problems in the template mappings. For instance, some articles of *Oil Fields by country* use the template *infobox_oil_field* (58%) whereas others use *infobox_oilfield* (23%), which have the same properties but were written in a different way.

This analysis also allows the detection of entities with different schemata but expressing distinct concepts within a same category. As an example, the category *Martial arts by type* contains 96 distinct templates (see Table 6). Although 63% of entities in this category instantiates the *Infobox boxer* template, this category does not contain only boxers: entities expressing other types of martial artists, events, teams, video games and championships can also be found.

Table 6 Overview of templates usage for each category. Total count of used templates, most used template, size and frequency of template most used

Category	*Templates*	*[l]Most Used template*	*Prop. (Most used)*	*Size (Most used)*
Alpine three-thousanders	4	infobox_mountain	0.9914	239
Formula One cars	3	infobox_racing_car	0.9807	108
French films	15	infobox_film	0.9877	36
HarperCollins books	15	infobox_book	0.9530	66
IOS games	17	infobox_video_game	0.8925	52
Martial arts by type	96	infobox_boxer	0.6255	31
Numbered minor planets	2	infobox_planet	0.9997	119
Oil companies of the USA	29	infobox_company	0.6458	84
Oil fields by country	21	infobox_oil_field	0.5859	79
Oil pipelines by country	5	infobox_pipeline	0.9444	43
Operas	22	infobox_opera	0.3720	25
Protein families	8	infobox_protein_family	0.9367	21
Skyscrapers between 100 and 149 metres	12	infobox_building	0.9260	182
Towns in Turkey	6	infobox_settlement	0.9870	424

5.4 Coverage of templates' attributes in infoboxes

In previous sections, we identify that the infobox template referenced by Wikipedia articles have a great influence in how infoboxes' articles are created. To verify the level of usage of the attributes in a template by the infoboxes, we calculate the proportion of infoboxes' attributes used from their mapped templates for each selected category using the Jaccard coefficient[14] as in equation (1).

$$jaccard\left(\vec{\delta},\vec{\gamma}\right)=\frac{\|\vec{\delta}\cap\vec{\gamma}\|}{\|\vec{\delta}\cup\vec{\gamma}\|} \quad (1)$$

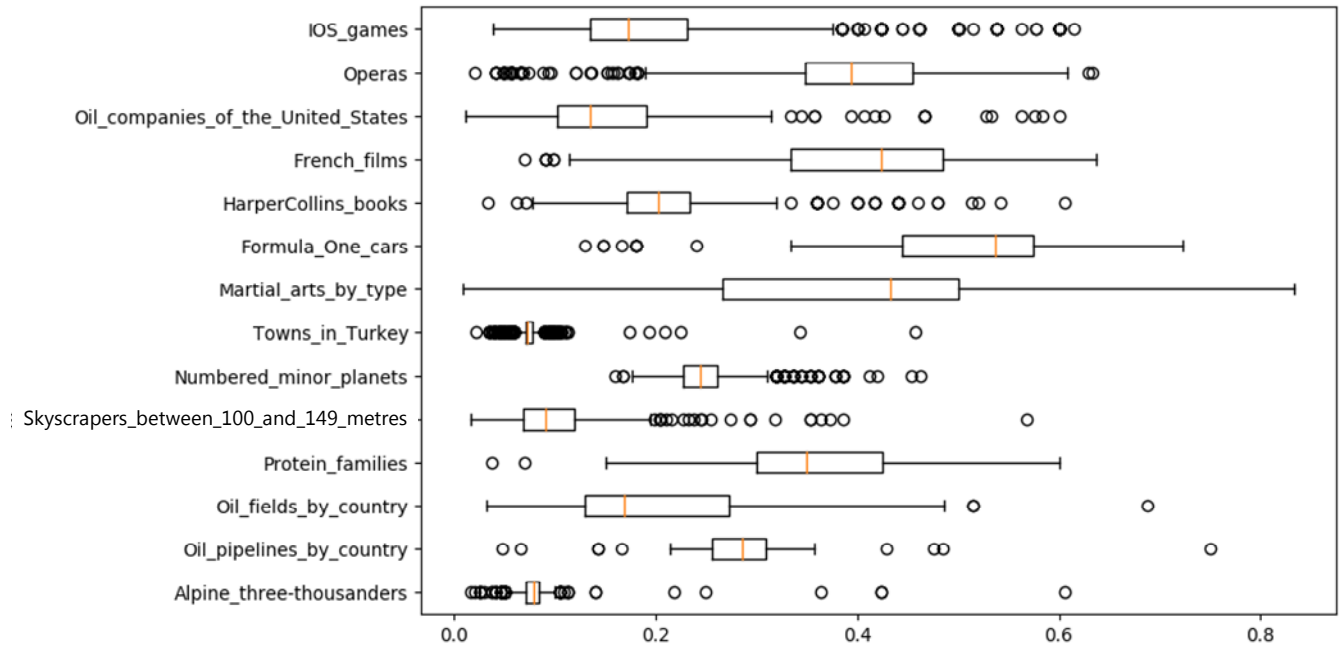
where:

$\vec{\delta}$, properties used on infobox instance;

$\vec{\gamma}$, suggested properties from mapped template;

Since, we remove duplicates to avoid inconsistent scores and $\vec{\delta}\subset\vec{\gamma}$, then $\|\vec{\delta}\cup\vec{\gamma}\|$ turns into $\|\vec{\gamma}\|$. Hence, this similarity coefficient can be read as a proportion or the ratio of actually used attributes from the suggested template. The result of this analysis can be seen in Figure 9. The numbers show that there is a great variation of the distribution of this proportion. While the categories *Formula One Cars*, *Martial arts by type*, and *French Films* present the highest values of median of proportion (close to 0.5), *Alpine three-thousanders*, *Towns in Turkey* and *Skyscrapers* are close to 0.

Figure 9 Distribution of proportions of used properties. The median is the quality index for category. Properties from suggested templates' schema used on internal schemata



When looking at the size of the most used suggested templates (see Table 6) and the actual size of infobox schemas for each category (see Figure 7), we notice what it seems to be a negative correlation: usually the most used template for a given category present a large number of properties while the median of infobox sizes for that same category is small. In other words, the larger the number of attributes on the templates, the lower is the infobox coverage. To verify this observation, we calculated the spearman correlation (Spearman, 1987) between the template coverage in the infobox's articles, equation (1), and its respective suggested template size. We perform this analysis using all recovered infoboxes from all analysed categories. The Spearman's coefficient is -0.7605 . Under a significance level of 0.05, the test presents p-value $< 2.2e^{-16}$, which indicates a high statistical significance level in terms of negative correlation between these variables. To illustrate this, a scatter plot between these two features are presented on Figure 10. An extreme case is the *Infobox_officeholder* template in the *Oil Companies* category, which has 1,400 properties but only nearly 0.03% of its attributes are actually used.

Figure 10 Scatter plot between template size and template coverage

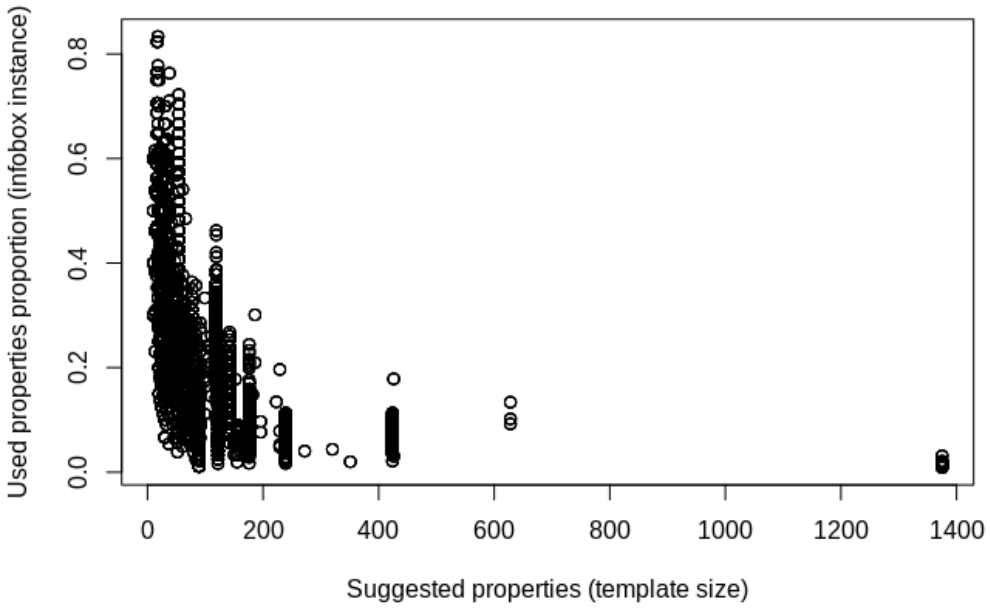
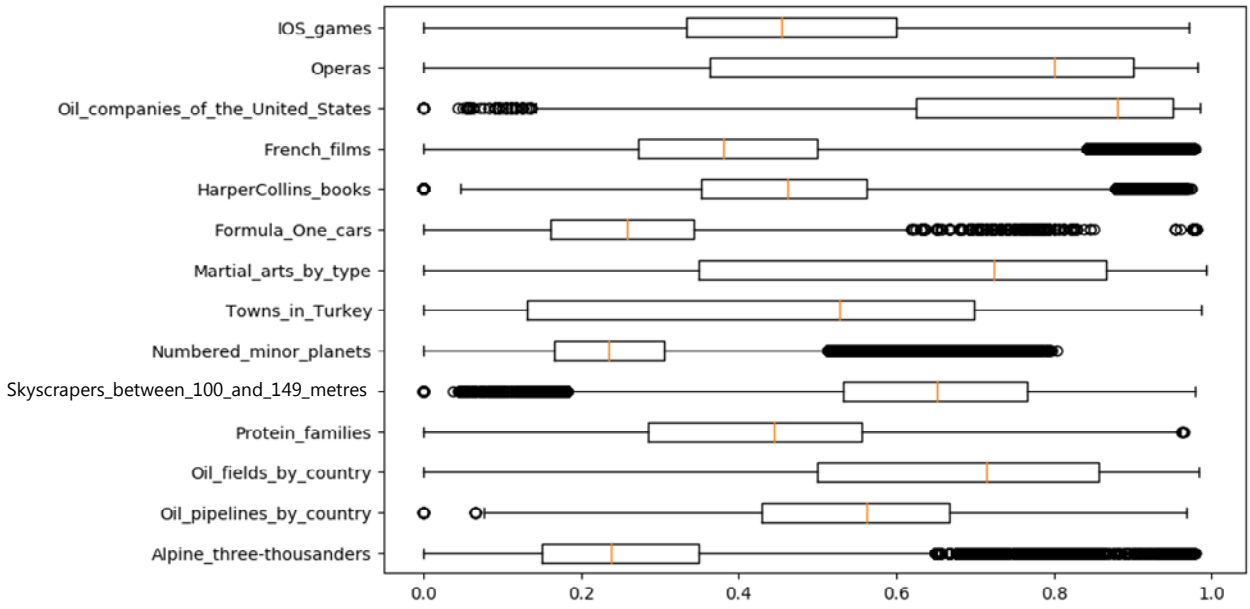


Figure 11 Infoboxes homogeneity. Jaccard similarity between infobox schemas



5.5 Homogeneity of infoboxes per category

In this section, we evaluate the homogeneity of the attributes of the infoboxes belonging to the same category. For that, we also used the Jaccard coefficient to measure the similarity between infoboxes from the same category based on their attributes. In Figure 11, we present the distribution of the similarity for the 14 categories. The plots show that in general there is not much homogeneity between infoboxes' properties belonging to the same category. For instance, the most heterogeneous categories are *Numbered minor planets* and *Alpine Three-thousanders*, both with median similarity around 0.2, despite the fact that on these categories only two and four templates respectively are used by the articles. In addition, for both categories there is a dominant infobox template: *Infobox_planet* in *Numbered minor planets* and *infobox_mountain* in *Alpine three-thousanders*, both with 99% of usage. The low Jaccard similarity in these two categories comes from the high number of properties present in their most popular templates: *Infobox_planet* (119) and *Infobox_mountain* (239), and the small median size of infoboxes in those categories: infoboxes of *Numbered Minor Planets* category have a median size of 29 attributes and infoboxes of *Alpine three-thousanders* category have a median size of 19 attributes.

In contrast, the most homogeneous categories are *Operas* and *Oil companies of the United States* with high median similarity between infoboxes: 0.8 and 0.9, respectively. This occurs even though their infoboxes refer a large number of template: 22 for *Operas* and 24 for *Oil companies of the United States*. This can be explained by the existence of shared properties across different templates. As an example, the category *Oil Companies of the United States* contains attributes such as owner, founder, location, architect and height which are present in many templates of this category.

We can conclude from these numbers that: (1) there is a great heterogeneity of infoboxes in terms of attributes within a same category and (2) the number of templates used by the infoboxes in a category has not much influence in its heterogeneity since different templates can share same attributes.

6 Summary of findings and prospects

The analysis performed here raises some questions and insights on how to improve quality of Wikipedia's structured data and how to leverage its use. Next, we highlight some findings including when necessary insights and questions to be explored:

1) *54% of the articles on Wikipedia have structured information*, what indicates that some effort is still required for automatic creation of Wikipedia Infoboxes, as for example the work done by Wu and Weld (2017); Wu et al. (2008) and Lange et al. (2010). Some strategies can be based on proposing new information extraction methods for mining this information from article's text or by integrating Wikipedia with external sources;
2) *the infoboxes' size follows a long-tailed distribution with median size of 13 attributes*;
3) *there is great occurrence of properties with different spelling but representing same semantic information*. Some work has been done regarding the multilingual matching of schemas (Nguyen et al., 2011; Rinser et al., 2013), however it is required some effort on reducing the noise of similar properties under the same language which could be done by semantic matching;
4) *there is a considerable amount of geographical and temporal information in infoboxes*. As an example, geospatial data can be used by visualisation tools while temporal data can be used for entities analysis as in Reznik and Shatalov (2016);
5) *there has been an effort to standardise the templates towards the "Infobox_" template*;
6) *the most popular infobox templates are also the ones representing generic concepts as settlement, person and organisation*. Investigating the influence of generic suggested templates on infobox quality would be helpful to guide the construction of more informative (complete) infoboxes;
7) *the number of attributes in suggested infobox templates has great influence in the infobox size distribution*;
8) looking at the ratio of suggested attributes that are actually present in infobox instances, we notice that: *the bigger the suggested infobox template, the lower the attributes coverage on infobox instances, and the smaller the suggested infobox template, the bigger the attributes coverage on infobox instances*;
9) *different templates can share same attributes*;
10) *there is a great heterogeneity of infoboxes' attributes within a same category*. These different attributes indicate a broad or too general category? Does it presents a relevant impact on the quality of Wikipedia articles, infoboxes and categories?

Beyond these findings this study can assist future works applying structured data from Wikipedia or other wikis. Owing this categories extent and scope our category approach leads to a better understanding of the coverage of schema diversity on a given domain. Infoboxes defined from a same template most times are distributed into different categories, which can represent slightly different domains. Hence, it is difficult to analyse schema diversity from an information domain by using only infoboxes defined from a same template, since templates present a predefined set of attributes.

Aligning categories hierarchy, infobox instances and templates we take a broad look on the diversity, organisation and inconsistencies of these schemas over different information domains. Also, most works using structured data from Wikipedia usually have to build their our data sets for training, test and validation to feed machine learning models. This step requires a specialised knowledge about data distribution in Wikipedia which requires time and effort either for understanding the data and to pre-process it. Beyond the

analysis to understand the data, our work makes available an indexing strategy for pre-processing and accessing this structured data.

7 Conclusions

Wikipedia presents a large space for structured data exploration. In this study, we performed a series of different analysis to help researchers understand Wikipedia data distribution and organisation.

To support the analysis we index Wikipedia structured data from its dump and DBpedia data sets. The pre-processing and indexing pipeline are publicly available as well as the final indexes for search, along with a basic search engine for querying and exploration of these indexes. As a result of this analysis we highlight some findings and point out some research questions related to the quantity and standardisation of Wikipedia structured data. Thus, these findings can be further explored for improving the quality of Wikipedia structured data. Last but not least the indexes we made available can be easily used for building data sets for training machine learning models.

This work can be further incremented with additional analysis, and explore other infobox templates beyond the ones starting with the *"Infobox_"* prefix. Also, it can be easily updated to a newer version of Wikipedia as soon as a new DBpedia dump is available. In order to remove the dependence on newer versions of DBpedia dumps, the categories hierarchy and infobox templates could be extracted directly from Wikipedia dump. Our next step for improving and extending the work presented here is to apply deep learning techniques for automatic measurement and classification of the quality of the defined infoboxes and articles in Wikipedia.

Acknowledgements

This work was supported by Fundação de Amparo a Ciência e Tecnologia do Estado de Pernambuco (FACEPE) under the funding grant No. IBPG-1172-1.03/16 and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) under the funding grant No. 88882.347588/2019-01.

References

Abbas, F., Malik, M.K., Rashid, M.U. and Zafar, R. (2016) 'Wikiqa' a question answering system on Wikipedia using freebase, dbpedia and infobox', *Proceedings of the 6th International Conference on Innovative Computing Technology (INTECH)*, pp.185–193.

Guda, B.P.R., Seelaboyina, S.B., Sarkar, S. and Mukherjee, A. (2020) 'NwQM: a neural quality assessment framework for Wikipedia', *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.8396–8406.

Heist, N. and Paulheim, H. (2019) 'Uncovering the semantics of Wikipedia categories', in Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M. and Gandon, F. (Eds): *The Semantic Web*, pp.219–236.

Hellmann, S., Lehmann, J., Auer, S. and Brümmer, M. (2013) 'Integrating nlp using linked data', *Proceedings of the 12th International Semantic Web Conference – Part II*, pp.98–113.

Jaccard, P. (1901) 'Étude comparative de la distribution florale dans une portion des alpes et des jura', *Bulletin de la Société Vaudoise des Sciences Naturelles*, Vol. 37, pp.547–579.

Kotlerman, L., Avital, Z., Dagan, I., Lotan, A. and Weintraub, O. (2011) 'A support tool for deriving domain taxonomies from Wikipedia', *Proceedings of the International Conferences on Recent Advances in Natural Language Processing*, pp.503–508.

Lalithsena, S., Perera, S., Kapanipathi, P. and Sheth, A. (2017) 'Domain-specific hierarchical subgraph extraction: a recommendation use case', *IEEE International Conferences on Big Data (Big Data)*, pp.666–675.

Lange, D., Böhm, C. and Naumann, F. (2010) 'Extracting structured information from Wikipedia articles to populate infoboxes', *Proceedings of the 19th ACM International Conferences on Information and Knowledge management*, pp.1661–1664.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P. and Auer, S. et al. (2015) 'Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia', *Semantic Web*, Vol. 6, No. 2, pp.167–195.

Lerner, J. and Lomi, A. (2018) 'Knowledge categorization affects popularity and quality of Wikipedia articles', *PLOS ONE*, Vol. 13, pp.1–22.

Lewoniewski, W. (2017) 'Completeness and reliability of Wikipedia infoboxes in various languages', in Abramowicz, W. (Ed.): *Business Information Systems Workshops*, Springer International Publishing, pp.295–305.

Lewoniewski, W. (2019) 'Measures for quality assessment of articles and infoboxes in multilingual Wikipedia', in Abramowicz, W. and Paschke, A. (Eds): *Business Information Systems Workshops*, Springer International Publishing, pp.619–633.

Mahdisoltani, F., Biega, J. and Suchanek, F. (2015) 'Yago3: a knowledge base from multilingual Wikipedias', *Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR)*.

Morales, A., Premtoon, V., Avery, C., Felshin, S. and Katz, B. (2016) 'Learning to answer questions from Wikipedia infoboxes', *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp.1930–1935.

Nguyen, T., Moreira, V., Nguyen, H., Nguyen, H. and Freire, J. (2011) 'Multilingual schema matching for Wikipedia infoboxes', *Proceedings of the VLDB Endowment*, Vol. 5, No. 2, pp.133–144.

Nguyen, T.H., Nguyen, H.D., Moreira, V. and Freire, J. (2012) 'Clustering wikipedia infoboxes to discover their types', *Proceedings of the 21st ACM International Conferences on Information and Knowledge Management*, pp.2134–2138.

Reznik, I. and Shatalov, V. (2016) 'Hidden revolution of human priorities: an analysis of biographical data from Wikipedia', *Journal of Informetrics*, Vol. 10, No. 1, pp.124–131.

Rinser, D., Lange, D. and Naumann, F. (2013) 'Cross-lingual entity matching and infobox alignment in Wikipedia', *Information Systems*, Vol. 38, No. 6, pp.887–907.

Rodriguez-Hernandez, I., Trillo-Lado, R. and Yus, R. (2020) 'Wikinfoboxer: a tool to create Wikipedia infoboxes using dbpedia', *XXI Jornadas de Ingeniera del Software y Bases de Datos*, Vol. 219, p.397.

Spearman, C. (1987) 'The proof and measurement of association between two things', *The American Journal of Psychology*, pp.441–471.

Suchanek, F.M., Kasneci, G. and Weikum, G. (2008) 'Yago: a large ontology from wikipedia and wordnet', *Journal of Web Semantics*, Vol. 6, No. 3, pp.203–217.

Titze, G., Bryl, V., Zirn, C. and Ponzetto, S.P. (2014) 'Dbpedia domains: augmenting dbpedia with domain information', *Proceedings of the 9th International Conferences on Language Resources and Evaluation (LREC'14)*, pp.1438–1442.

Vivaldi, J. and Rodrguez, H. (2010) 'Finding domain terms using wikipedia', *Proceedings of the 7th International Conferences on Language Resources and Evaluation (LREC'10)*, 2010.

Wecel, K. and Lewoniewski, W. (2015) 'Modelling the quality of attributes in wikipedia infoboxes', Witold Abramowicz, editor, *Business Information Systems Workshops*, Springer International Publishing, pp.308–320.

Wu, F. and Weld, D.S. (2017) 'Autonomously semantifying wikipedia', *Proceedings of the 16th ACM International Conferences on Information and Knowledge Management*, pp.41–50.

Wu, F., Hoffmann, R. and Weld, D.S. (2008) 'Information extraction from wikipedia: Moving down the long tail', *Proceedings of the 14th International Conferences on Knowledge Discovery and Data Mining, on Knowledge Discovery and Data Mining*, pp.731–739.

Zhang, K., Xiao, Y., Tong, H., Wang, H. and Wang, W. (2014) 'Wiicluster: a platform for wikipedia infobox generation', *Proceedings of the 23rd ACM International Conferences on Information and Knowledge Management*, pp.2033–2035.

Notes

1 https://en.wikipedia.org/wiki/Template:Grading_scheme

2 http://wikidata.dbpedia.org/develop/datasets/dbpedia-version-2016-10

3 https://www.w3.org/2009/08/skos-reference/skos.html

4 https://en.wikipedia.org/wiki/Category:Infobox_templates

5 https://en.wikipedia.org/wiki/Help:Infobox

6 https://mwparserfromhell.readthedocs.io/en/latest/

7 https://lucene.apache.org/core/

8 https://cin.ufpe.br/~jms5/complete-infobox-index/

9 https://github.com/guardiaum/InfoboxIndexSearch

10 https://en.wikipedia.org/wiki/Al_Jazeera_English

11 https://en.wikipedia.org/wiki/Template:Infobox_television_channel

12 https://en.wikipedia.org/wiki/Template:Infobox_football_biography

13 https://en.wikipedia.org/wiki/Template:Infobox_settlement

14 To improve the matching between attributes, we normalise the property names through lowercasing, removing underlines, dashes, spaces and duplicates.