Evolution of disease-related content on collaborative platforms: a longitudinal analysis

Dr. G. Moretti[1], Dr. F. Romano[1], Dr. L. Esposito[1], Dr. P. De Luca[1]*

[1] Department of Internal Medicine and Clinical Research, University of Naples Federico II, Naples, Italy

Abstract

Wikipedia, also known as "The Free Encyclopaedia", is one of the largest online repositories of biomedical information in the world, and is nowadays increasingly been used by medical researchers and health professionals alike. In spite of its rising popularity, little attention has been devoted to the understanding of how such medical information is organised, and especially how it evolves through time. We here present an analysis aimed at characterising such evolution, with a focus on the effects that such dynamic may have on an automated knowledge extraction process. For that, we start from a data set comprising a large number of snapshots of Wikipedia's disease articles, and the corresponding diagnostic elements as provided by the DISNET project (disnet.ctb.upm.es). We then track and analyse how different metrics evolve through time, such as the total article length or the number of medical terms and references. Results highlight some expected facts, as for instance that most articles increase their content through time; and that hot topics, as Alzheimer's disease, attract the highest number of editions and views. On the other hand, relevant behaviours are observed for less well-known diseases, including abrupt changes in the text and the concentration of contributions in a handful of editors. These results stress the importance of using correctly filtered and up-to-date datasets, and more general of considering the temporal evolution of the information in Wikipedia.

Keywords: wikipedia disease, diagnostic knowledge, information retrieval, change knowledge, wikipedia evolution, medical content

1 Introduction

Everything is constantly changing (Boudoulas, Triposkiadis, Stefanadis, & Boudoulas, 2017) with an ever increasing frequency (Densen, 2011), and medicine is no exception to this, with new advances and solutions relentlessly appearing. This is partly thanks to the integration and joint evolution of medicine and technology, which have produced a large amount of medical solutions that have improved the health of the world's population (Fernandez-Moure, 2016). Each significant advance implies an increase in medical knowledge. In the middle of the last century, according to the Fuller's study, the time in which knowledge doubled (Knowledge Doubling Curve (Fuller, 1982)) was 50 years. In medicine, this knowledge doubling time was already down to 3.5 years in 2010, thanks to new technologies and tools, but it's estimated to reach a record 73 days by 2020 (Densen, 2011). All this continuous flow of medical discoveries about the origin of diseases, symptoms, cures and treatments is supposing a challenge for researchers and health professionals, as such knowledge is required both to improve current medical practices, and both to generate new knowledge.

The Internet has undoubtedly been the platform that has boosted the globalisation of information. Internet has further made feasible many public and private biomedical information sources that are frequently used for research purposes, such as UMLS (Bodenreider, 2004), PubMed (Lindberg, 2000), OMIM (Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005), DisGeNet (Piñero et al., 2017), Human Phenotype Ontology (HPO) (Köhler et al., 2017; Robinson et al., 2008), Disease Ontology (DO) (Schriml et al., 2012), MeSH (Lipscomb, 2000), MalaCards (Espe, 2018; Rappaport et al., 2013, 2014), GeneCard (Safran et al., 2002), Diseases Database (DD)[1], DISEASES (Pletscher-Frankild, Pallejà, Tsafou, Binder, & Jensen, 2015), MayoClinic ("Mayo Clinic," 2019), MedlinePlus (Miller, Lacroix, & Backus, 2000), or Wikipedia (Thompson & Hanley, 2017). Each source has its own idiosyncrasies, both in the way information is collected and organised, and on the limitations imposed on its use – such that, in many cases, their information is not

[1] http://www.diseasesdatabase.com

freely accessible, or is partially available (Wang, Pourang, & Burrall, 2019). It is also important to highlight that the constant growth of medical knowledge and, to the same proportion, the increase in the number of available biomedical repositories, complicate the tasks of performing searches and surfing through them; hence the new need for automatic medical data recollection (Dias, Oliveira, Vicente, & Martín-Sánchez, 2005; Lopes & Oliveira, 2013; Murray, 2019; Oliveira et al., 2004). In this context, public elements (public medical sources) play a major role in the dissemination of medical knowledge, as they promote a philosophy of collaboration, growth and continuous improvement.

Wikipedia, also known as "The Free Encyclopaedia", is one of the most visited websites in the world, and the largest freely accessible encyclopaedia covering several fields of human knowledge. It is further emerging as one of the major providers of medical knowledge (G. Lagunes García et al., 2019), both for the general population, and for the scientific community and health care professionals (Allahwala, Nadkarni, & Sebaratnam, 2013; Hodson, 2015; Tucker, 2014). Despite having some detractors, there are many research works that report the great potential and quality of the information there gathered since its launch in 2001 (Al Tamime, 2017; Allahwala et al., 2013; Azer, 2015; Cohen, 2013; Fairchild, Del Valle, De Silva, & Segre, 2015; Fairchild et al., 2015; Friedlin & McDonald, 2010; G. Lagunes García et al., 2019; Goslin & Hofmann, 2018; Shafee et al., 2017; Valle et al., 2018). Wikipedia responds not only to the need of having a platform for integrating the knowledge coming from different places, but also to the necessity of having a vehicle in which the sum of all human knowledge is available at first-hand (Jiang, Bai, Zhang, & Hu, 2017; Mehdi, Okoli, Mesgari, Nielsen, & Lanamäki, 2017; Mesgari, Okoli, Mehdi, Nielsen, & Lanamäki, 2015). Such is the impact that Wikipedia is having on science, that is influencing one in every eight hundred words appearing in scientific articles (Thompson & Hanley, 2017). The relevance of this controversial and popular source in the medical field is undeniable; and its impact is only expected to increase, having the potential to be one of the largest providers and disseminators of biomedical knowledge.

Given the volume of meaningful and useful medical knowledge contained in Wikipedia (Al Tamime, 2017; Allahwala et al., 2013; Azer, 2015; Cohen, 2013; Fairchild et al., 2015; Friedlin & McDonald, 2010; G. Lagunes García et al., 2019; Mesgari et al., 2015; Shafee et al., 2017; Thompson & Hanley, 2017; Tucker, 2014; Valle et al., 2018), and its non-completed nature, little attention has been devoted to its evolution. Such analysis would shed light on several aspects of this encyclopaedia, as the appearance of trends, styles, or editing behaviours. Additionally, it would provide the opportunity to follow the trail that human medical knowledge has left its way, or, in other words, to understand how medical knowledge evolves. Finally, the results of such analyses could be used to improve automatic knowledge extraction processes.

In line with this idea, this research work aims at analysing how the medical content of those articles catalogued as diseases in the English version of Wikipedia has evolved. Towards this objective, we consider the history of the revisions of each article, and specifically focus on three elements: **i)** the evolution of the number of characters, **ii)** of references, and **iii)** of medical terms. We further study the evolution of these metrics in an independent way, and through the detection of correlations between them. Results highlight, among others, the unexpected behaviours of minor articles, whose evolution is characterised by sudden and major changes, and whose editing is concentrated in the hands of few editors.

Beyond the introduction, this article is organised as follows. Section 2 discusses the background of the study, and specifically how it leverages on some previously presented results. Section 3 then describes the methodological and technical aspects of the analysis, including the considered data source. Finally, Section 4 details the obtained results, and Section 5 draws some conclusions.

2 Background

This section describes the main data source used in the analyses here presented, as provided by the DISNET project; it also describes previous research work on the same data and its limitations, which support the development of the new analyses.

2.1 DISNET

The analyses here reported leverage on the DISNET project[2], whose ultimate goal is to improve the understanding of diseases through their representation as a complex multi-layer network. The first stage of this project encompassed the development of a Web platform that, using a novel approach, is capable of obtaining disease information from several public sources, including (Gerardo Lagunes García et al., 2018). Of relevance here is the capability of retrieving information relative to the phenotypic manifestations of each one of the articles of the Wikipedia in English catalogued as diseases, as contained in those sections

[2] disnet.ctb.upm.es

discussing these elements. The extraction process is executed automatically every 15 days, being currently in execution since the beginning of its activity, February 1st, 2018. Each execution of the process involves, firstly, the generation of a snapshot of the text including phenotypic knowledge; and secondly, the execution of an NLP process to extract a list of medical terms referring to signs and symptoms, which DISNET call Diagnostic Knowledge Elements (**DKEs**). The resulting data set thus includes relationships between diseases and related DKEs integrated from different sources – note that the system also retrieves information from other data sources, here not considered. All DISNET data are freely available online through an easy to use API, whose documentation is available online at the following link[3]. To illustrate, the interested researcher may use DISNET queries for obtaining, for instance, the phenotypic manifestations associated to Influenza on February 1st, 2018 or February 1st, 2019, both on Wikipedia and PubMed.

It is important to emphasise that the NLP pipeline used by DISNET is fully responsible for the search and detection of relevant DKEs; errors, like the missed detection of a DKE, would propagate in subsequent studies using those data.

2.2 Previous work

The work here presented leverages on results previously obtained under the umbrella of the DISNET project, aimed at describing the evolution of the elements of Wikipedia disease articles that are used to guide the diagnostic process, i.e. the evolution of DKEs (Lagunes García et al., 2019). More specifically, these results include the temporal evolution of the number of DKEs during one year, from the February 1st, 2018 to February 1st, 2019. The sections there considered are: "Signs and symptoms", "signs and symptoms", "Symptoms and causes", "Signs", "Symptoms", "Causes", "Cause", "Diagnosis", "Diagnostic", "Causes of injury", "Diagnostic approach", "Presentation", "Symptoms of ...", "Causes of ...". and infobox. Table 1 shows the evolution by snapshot of the number of DKEs and of other elements found in these sections, suggesting a global upward trend.

Table 1. Evolution of the information contained in the snapshots: number of articles retrieved by DBpedia as diseases (DBpDis), number of Wikipedia articles that contain DKE applying TVP validation (WRDArt), of the number of DKE found by MetaMap (not applying TVP validation but removing duplicates) (WRawDF), number of texts (WTxt), number of semantic types found (again: before applying TVP) (WST), number of external codes found in Wikipedia (WExCd), number of external sources found in Wikipedia (WExtSrc) and number of links found in the texts (WLink). Source: (Lagunes García et al., 2019).

Snapshot	DBpDis	WRDArt	WRawDF	WTxt	WST	WExCd	WExtSrc	WLink
2018-02-01	8,161	3,625	13,332	30,932	17	19,203	61	148,640
2018-02-15	8,161	3,631	13,356	31,009	17	19,161	60	149,073
2018-03-01	8,161	3,636	13,393	31,199	17	19,116	60	149,799
2018-03-15	8,161	3,644	13,431	31,463	17	19,102	60	151,065
2018-04-01	9,857	3,841	13,598	32,956	17	19,302	62	157,585
2018-04-15	9,858	3,846	13,608	33,016	17	19,289	62	158,107
2018-05-01	9,858	3,860	13,630	33,153	17	19,272	62	158,910
2018-05-15	9,858	3,871	13,692	33,246	17	19,270	62	159,262
2018-06-01	9,858	4,047	13,853	34,376	17	19,246	58	163,710
2018-06-15	9,858	4,061	14,169	34,584	18	19,224	58	164,709
2018-07-01	9,858	4,062	13,980	34,953	17	19,209	59	165,732
2018-07-15	9,858	4,069	13,991	35,044	17	19,207	58	166,127
2018-08-01	9,858	4,083	14,013	35,153	17	19,202	58	166,814
2018-08-15	9,858	4,087	14,019	35,247	17	19,201	58	167,300
2018-09-01	9,858	4,090	14,022	35,296	17	19,333	58	167,584
2018-09-15	9,858	4,091	14,022	35,293	17	19,339	58	167,688

[3] disnet.ctb.upm.es/apis/disnet

Snapshot	DBpDis	WRDArt	WRawDF	WTxt	WST	WExCd	WExtSrc	WLink
2018-10-01	9,858	4,093	14,032	35,343	17	19,351	58	167,928
2018-10-15	9,858	4,097	14,050	35,449	17	19,360	58	168,497
2018-11-01	9,858	4,101	14,064	35,522	17	19,360	58	168,972
2018-11-15	9,858	4,106	14,089	35,657	17	19,366	58	169,745
2018-12-01	9,858	4,111	14,140	35,778	17	19,355	59	170,344
2018-12-15	9,858	4,119	14,159	35,949	17	19,365	59	171,203
2019-01-01	9,858	4,128	14,182	36,111	17	19,365	59	172,082
2019-01-15	9,858	4,139	14,185	36,192	17	19,366	59	172,456
2019-02-01	11,084	4,692	14,722	40,913	17	20,237	59	193,860

This previous analysis also covered a comparison of the variation in the number of diseases between each snapshot. Table 2 shows the increase in the number of diseases, the overlap with respect to the previous snapshot, and the variation as a percentage.

Table 2. Comparison between disease snapshots. Source: (Lagunes García et al., 2019).

Snapshot	No. Disease	% Overlapping	% Disease Variation
2018-02-01	-	-	-
2018-02-15	+6	99.83	2.12
2018-03-01	+8	99.78	3.20
2018-03-15	+15	99.59	2.98
2018-04-01	+203	94.71	3.66
2018-04-15	+10	99.74	2.55
2018-05-01	+21	99.46	3.10
2018-05-15	+13	99.66	12.34
2018-06-01	+179	95.58	7.45
2018-06-15	+18	99.56	44.40
2018-07-01	+10	99.75	45.26
2018-07-15	9	99.78	2.44
2018-08-01	+19	99.53	2.61
2018-08-15	+6	99.85	2.67
2018-09-01	+7	99.83	1.86
2018-09-15	+5	99.88	1.84
2018-10-01	+4	99.90	2.18
2018-10-15	+9	99.78	1.81
2018-11-01	+12	99.71	2.93
2018-11-15	+9	99.78	2.29
2018-12-01	+10	99.76	3.24
2018-12-15	+9	99.78	2.31
2019-01-01	+12	99.71	3.04
2019-01-15	+11	99.73	1.89
2019-02-01	+558	88.12	3.29

2.3 Limitations of the previous approach

In summary, the previous work was centred on the analysis of the evolution of Wikipedia disease articles from the point of view of their most valuable elements, i.e. DKEs. Still, medical terms are not the only information sources; and sections beyond those involving phenotypic manifestations may also be of interest.

To illustrate the limitations of this approach, we here focus on the evolution of the article for the Otic polyp disease. Fig. 1 reports two screenshots, respectively corresponding to February 1st, 2019 and February 15th, 2019. It can be appreciated that the text is the same in both cases; still, DISNET detects an important increase in the number of DKEs between the two snapshots. The reason resides in how information is organised: while in the first snapshot only one section ("Signs and symptoms") is recognised as relevant, and the information of "Imaging findings", "Pathology findings" and "Differential diagnoses" is disregarded, these latter three have been grouped into a single section in the second snapshot, i.e. "Diagnosis", which is indeed processed by DISNET. Therefore, a restructuring of the content, which is only moving information, may be wrongly interpreted by DISNET as an increase in the quantity of available information.

We here thus follow a complementary approach, focusing on the whole articles and on the dynamics of their textual content, as is detailed in the next Section.

3 Materials and Methods

This section describes the workflow of the process retrieving historical information from disease articles in Wikipedia; it also describes the dataset resulting from the execution of our historical extraction process; as well as the approach proposed for performing the data analysis.

3.1 Procedure to retrieve historical information from disease articles on Wikipedia

As previously introduced, the workflow here developed leverages on DISNET for input information, and specifically for each disease on: the name of the disease, i.e. the title of the Wikipedia article; the link to the Wikipedia article; and the list of snapshots with associated access time stamps.

With the information described above it is possible to access the historical content of Wikipedia articles through revisions, which are updates or editions that have been performed on a specific article, on a specific date, and by a given user. Wikipedia identifies the articles by their title or page id; in a similar way, previous versions of each document are identified by a unique revision id. To illustrate, the following link[4] leads to the Otic polyp disease article; the link[5] leads to its revision history; and the link[6] leads to a specific revision of the article, in this case to the revision made on January 10th, 2019. Fig. 2 shows how the previously described links are displayed in Wikipedia. This example highlights the two key elements for the historical information retrieval procedure: the title of the article and the id of the revision (oldid). Still, DISNET only returns the former, but not the complete list of revision ids; an additional process had then to be developed.

Two possible approaches for accessing the historical content of the articles were evaluated: a Web scraping process similar to the one developed by DISNET on one hand; and the use of the English Wikipedia API[7], an endpoint of the MediaWiki action API[8], both from the Wikimedia Foundation, Inc. family[9]. The latter option has been selected as, being an API, it offers a standard language for queries. Additionally, the API also allows to develop a great variety of requests, whose output is already structured in a way optimised for further processing. By comparison, reaching the same results with a custom Web crawler would require an important development effort.

[4] https://en.wikipedia.org/wiki/Otic_polyp
[5] https://en.wikipedia.org/w/index.php?title=Otic_polyp&action=history
[6] https://en.wikipedia.org/w/index.php?title=Otic_polyp&oldid=877794912
[7] https://en.wikipedia.org/w/api.php
[8] https://www.mediawiki.org/wiki/API:Main_page/en
[9] https://es.wikipedia.org/wiki/Fundaci%C3%B3n_Wikimedia

Fig. 1. Comparison between two revisions of the Wikipedia article on Otic polyp. a) Revision that matches the DISNET snapshot of February 1st, 2019; b) Revision that matches the DISNET snapshot of February 15th, 2019

Two different queries are required to obtain the historical data of the articles.

The first, **WQone**, allows to retrieve the basic information of the revision of a specific article, where the returned revision is the last one before the date specified in the query. The configuration of the request, shown in Fig. 3, consists of four basic elements. Firstly, it is necessary to specify that the request is aimed at obtaining data: the action[10] parameter is then set to query. Secondly, the revisions value of the prop parameter indicates that the query focuses on information about revisions. Thirdly, four parameters are used to retrieve only data of relevance in this study:

1) rvprop, which specifies the list of attributes to be return: {(ids, the revision id), (flags, the revision flags "minor"), (timestamp, the timestamp of the revision), (userid, the user id of the revision creator), (user, user that made the revision), (size, the length (bytes) of the revision) and (comment, comment by the user for the revision)}.

2) rvstart, defining the revision date; note that, if nothing is here specified, the query will return information about the revision defined by the rvdir parameter.

3) rvdir, which specifies how revisions should be sorted. In this case the older value is used, in order to get the oldest revision before the target date.

4) rvlimit, i.e. the number of revisions to return, in this case one.

[10] https://en.wikipedia.org/w/api.php?action=help&modules=main

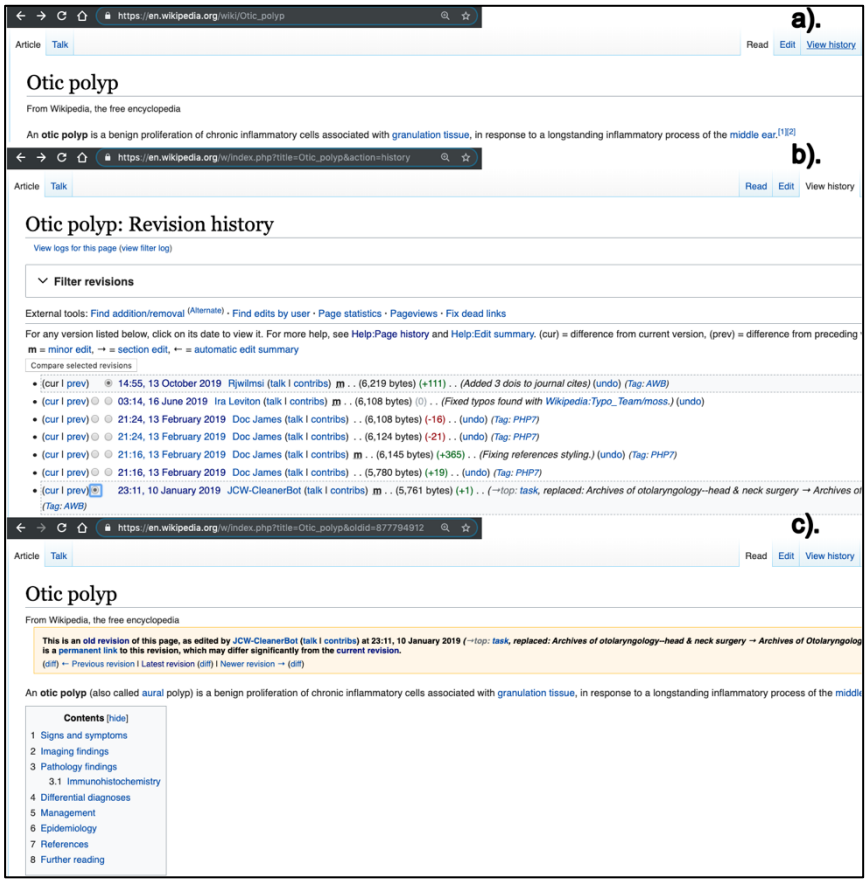

Fig. 2. Track to look at a revision of the Wikipedia article on Otic polyp. a) Current version of this Wikipedia disease article, as at September 2019; b) current revision history; and c) content of one of its specific revisions (January 10th, 2019).

The last set of parameters of WQone are titles, which specifies the name of the Wikipedia article (name of the disease) to be searched; redirects, which allows to follow an article through changes in names and redirections (a very common case in Wikipedia); and finally the format parameter, which allows to indicate the response format, in our case json. For more information about revision queries in general, and the type of query action=query in particular, the interested reader may refer to the documents respectively available at[11] and[12].

It is true that there is the option to recover the textual content of the revision directly with WQone, but it is a format that is not structured and neither has the option to retrieve some elements of the document, such as the list of sections, images, links, among other elements, nor is the option to format the text to HTML, a format that has several methods of undermining its content. Therefore, a second query (WQtwo) was developed.

Even though WQone has an option to retrieve the full text of a revision, this is not returned structured, nor it is possible to obtain specific elements like list of sections, images, or links. Towards this aim, a second query (**WQtwo**) has been developed, which retrieves all textual content in HTML format and the list of sections of a given revision. For this, it firstly uses parse as a value for the action parameter, to analyse the content of a page and obtain the output. Secondly, the prop parameter allows to lists the elements to be recovered from the document, in the present case “sections|text”. Thirdly, format=json indicates the response format of the query; and finally, oldid specifies the id of the revision. An example of the use of WQtwo is reported in Fig. 4.

[11] https://www.mediawiki.org/wiki/API:Revisions
[12] https://www.mediawiki.org/wiki/API:Query

Request: https://en.wikipedia.org/w/api.php?action=query&prop=revisions&format=json&rvprop=ids|flags|timestamp|userid|user|size|comment&rvstart=2019-01-15T00:00:00Z&rvdir=older&rvlimit=1&redirects&titles=Otic%20polyp

Parameter breakdown:

https://en.wikipedia.org/w/api.php

```
action    = query
prop      = revisions
format    = json
rvprop    = ids|flags|timestamp|userid|user|size|comment
rvstart   = 2019-01-15T00:00:00Z
rvdir     = older
rvlimit   = 1
redirects =
titles    = Otic polyp
```


```
{
    "continue": {
        "rvcontinue": "20160527211659|722399173",
        "continue": "||"
    },
    "query": {
        "pages": [
            {
                "pageid": 35510049,
                "ns": 0,
                "title": "Otic polyp",
                "revisions": [
                    {
                        "revid": 877794912,
                        "parentid": 722399173,
                        "minor": true,
                        "user": "JCW-CleanerBot",
                        "userid": 31737083,
                        "timestamp": "2019-01-10T23:11:39Z",
                        "size": 5761,
                        "comment": "/* top */[[User:JCW-
CleanerBot#Logic|task]], replaced: Archives of
otolaryngology--head & neck surgery → Archives of
Otolaryngology—Head & Neck Surgery"
                    }
                ]
            }
        ]
    }
}
```


Fig. 3. Wikipedia query 1 (WQone): Get revision basic information by disease name (article title) and snapshot

Fig. 5 depicts the complete workflow of our historical information extraction procedure. It starts with the module in charge of retrieving data from DISNET, specifically the list of diseases (name and DISNET internal id) that were collected during 18 months, along with the respective list of snapshots (dates in which the system retrieved the data). Next, a different module consumes the Wikipedia API through the execution of WQone for each disease, and WQtwo for each snapshot. It also extracts some relevant metrics from each snapshot, i.e. the number of characters and of references; and finally saves all results in JSON files. As a final step, the analysis module is in charge of filtering the obtained data, applying a descriptive analysis (which is described in Section 3.2) and visually representing the results. The resulting collection of JSON files is available online at the following link[13]. We also make available the code of our historical information extraction procedure in the following repository[14].

[13] https://midas.ctb.upm.es/gitlab/disnet/ipm2019/blob/master/JSON_files.zip

[14] https://midas.ctb.upm.es/gitlab/disnet/ipm2019/blob/master/retrieval_procedure_code/rhdwiki.zip

Fig. 4. Wikipedia query 2 (WQtwo). Get revision full textual information and section list

At this point, it is worth discussing two specific situations that cannot correctly be processed by the system. First of all, Wikipedia offers its users both the possibility of deleting an entire article with all its revisions, and of selectively deleting one or more revisions of an article. It is clear that, in this case, the developed extraction process is not able to recover any kind of historical information. An example of this can be found in the Wikipedia article Pancreatic mucinous cystic neoplasm. While information was correctly recovered from February 1st, 2018 to November 1st, 2018, starting from November 2019 the article was no longer available due to "Unambiguous copyright infringement"[15]. The second special situation is consequence of the evolving nature of both DISNET and Wikipedia, as both of them have been expanding their respective lists of diseases – the former thanks to internal improvements, and the latter as a consequence of additions. Consequently, not all diseases were present throughout the 18 months considered by this study, and, for some of them, snapshots may be missing

In order to solve these inconsistencies, a filter was applied, according to which only those articles that have data in each of the 36 snapshots have been considered. The result of the filtering process generated three new datasets with information on 3,508 diseases that is described in section 2.2.

[15] https://en.wikipedia.org/wiki/Special:Log?type=delete&user=&page=Pancreatic_mucinous_cystic_neoplasm&wpdate=&tagfilter=&subtype=

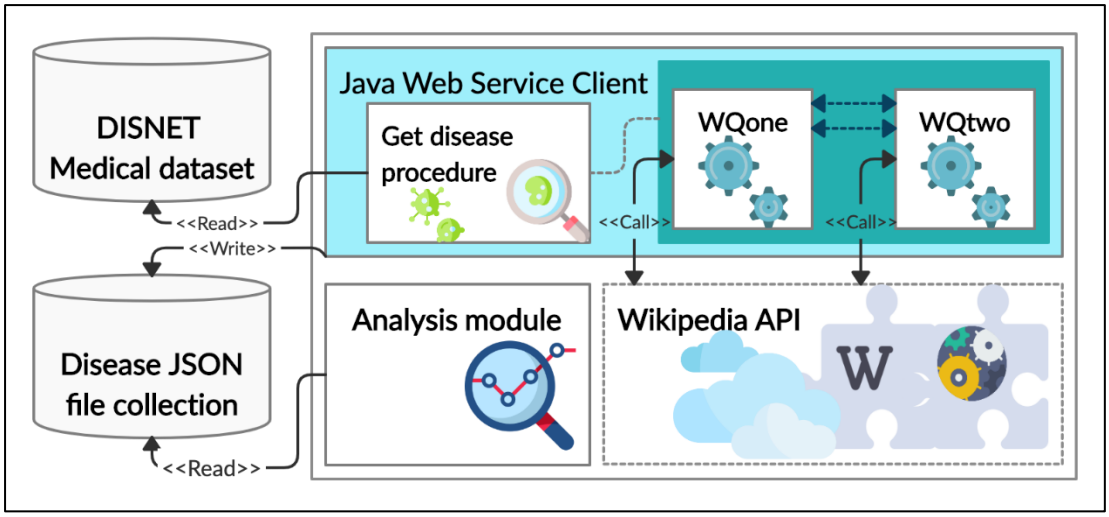

Fig. 5. Procedure to retrieve historical information from disease articles on Wikipedia workflow. (Icon made by Freepik from www.flaticon.com).

3.2 Analysis design

In order to analyse how the content of disease Wikipedia articles has evolved, we have focused on the measurement and observation of three variables: the number of characters in the article, the number of references, and the number of DKEs. As further detailed in Section 4, the analysis both focuses on the individual metrics, as well as on the relationships between pairs of them.

The number of characters is a measure that allows us to easily recognise when an article has been updated or modified. Note that modifications can involve adding more text with new information, removal of text in depuration mode to achieve a more synthetic version, and, finally, changes in the way ideas are expressed. If the number of characters focuses on the quantity, references and DKEs are more related to quality: the former by quantifying the level of leverage on scientific literature, the latter the density of medical concepts. Note that the relation between these three metrics is far from being trivial, as an increment in one of them does not necessarily imply an increment in the other two.

As a final note, all the results obtained by this analysis process have been synthesised and published online, to facilitate future studies. These are: **a) the variation in the number of characters per disease**[16], **b) the variation in the number of references per disease**[17] and **c) the variation in the number of medical terms per disease**[18]. Each data set is structured as a matrix, in which each row is a disease and each column is the number of characters, references or DKEs (according to the dataset) in each of the 36 DISNET snapshots.

4 Results

In this section we report several analyses based on the incremental evolution of the three previously described metrics, i.e. number of characters, terms and references in each page.

Fig. 6, left panel, firstly represents a histogram of the number of snapshots as a function of the percentage of change in the number of characters (black line), terms (blue line) and references (red line), with respect to the previous snapshot of the same disease. In other words, given two consecutive snapshots at time t and $t+1$, we calculate the relative change as $v_m(t+1) = 100.0 \cdot (m_{t+1} - m_t) / m_t$, m being the metric under analysis. Two interesting ideas can be drawn. First of all, the three distributions are characterised by a long tail, as for instance it is possible to find a significant number of snapshots that have modifications in their metrics of 40% or more. This is to be expected in the case of references and terms, as it is possible that content is added to snapshots that initially only included very short texts, thus with only a handful of terms. Yet, the same behaviour is observed for the number of characters; modifications can thus

[16] https://midas.ctb.upm.es/gitlab/disnet/ipm2019/blob/master/wikipedia_diseases_characters.csv
[17] https://midas.ctb.upm.es/gitlab/disnet/ipm2019/blob/master/wikipedia_diseases_dkes.csv
[18] https://midas.ctb.upm.es/gitlab/disnet/ipm2019/blob/master/wikipedia_diseases_references.csv

frequently go beyond simple grammar corrections but may include profound text enhancements. This is further confirmed in the right panel of Fig. 6, which represents the cumulative probability distribution of the fraction of snapshots that have witnessed as given change in the three metrics. In mathematical terms, this probability is defined as $P_m(v) = P(m \geq v)$, where m is the considered metric, and v the variation being evaluated. It can be appreciated that, especially for the number of references, most of the changes suppose a variation of more than a 50%.

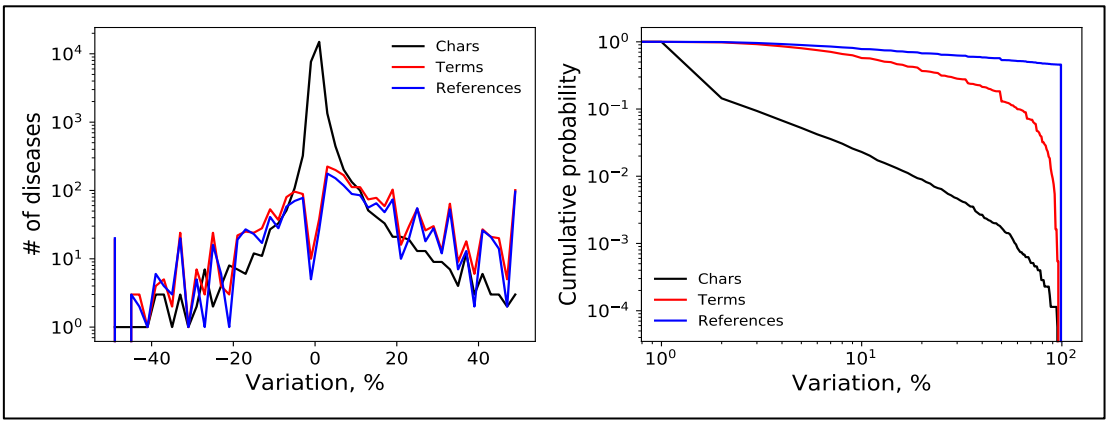

Fig. 6. Evolution of Wikipedia's snapshots. (Left) Probability distribution of the variation of the number of characters (black line), DKEs (red line) and references (blue line) in two consecutive snapshots, expressed as a percentage. (Right) Cumulative probability distribution of the same three metrics.

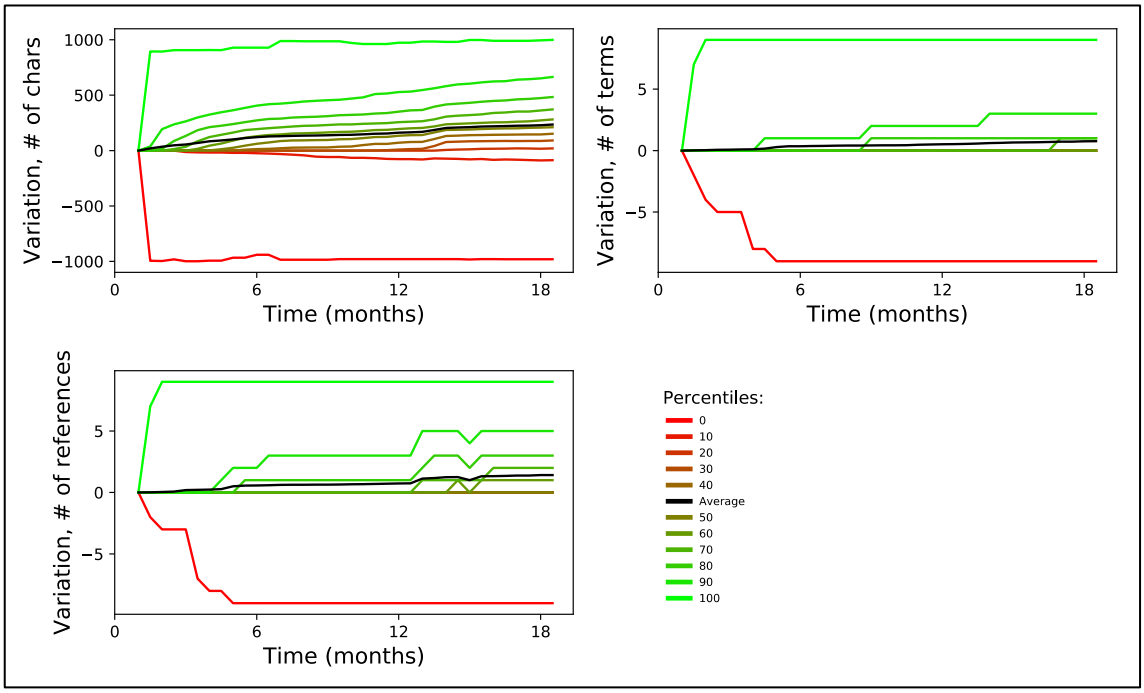

Fig. 7. Evolution of the number of characters, DKEs and references, throughout all the snapshots. Each colour share represent a decile of the distribution (see right legend). All values are normalised such that the initial value is zero.

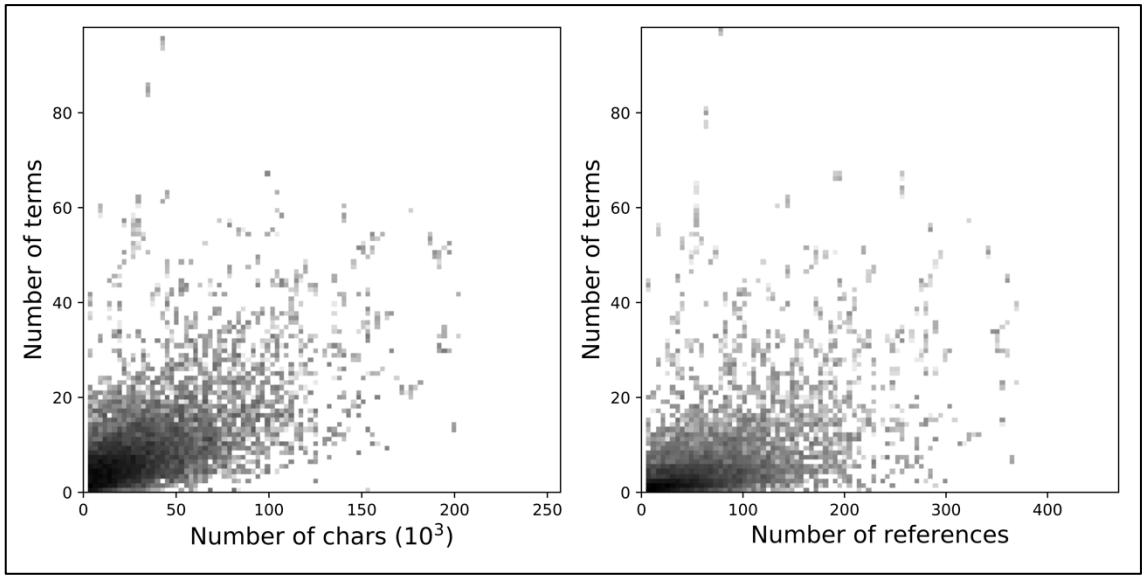

Fig. 8. Relationships between the three considered metric. The left (respectively, right) density map depicts the number of snapshots with a given number of characters and terms (references and terms). Darker shades indicate higher numbers of snapshots

Coming back to the left panel of Fig. 6, while most modifications imply an increase in the content, it is interesting to note that some of them display a substantial reduction. Specifically, a reduction is observed in the 32.17% of snapshots in the case of the number of characters, 25.16% in the case of terms, and 15.60% in the case of references. This may be due to splits, e.g. when a single disease is then splitter into two separate concepts; but also, to major restructuring in which redundant information is deleted.

Table 3. Disease articles with highest and lowest character average

Metric	Disease article	Min and max characters	Number of changes
1st highest character average	Drowning	161,160 - 167,190	21
2nd highest character average	HPV-positive oropharyngeal cancer	150,073 - 162,228	17
3rd highest character average	Alzheimer's disease	152,351 - 166,992	28
4th highest character average	Major depressive disorder	150,271 - 159,412	28
5th highest character average	HIV/AIDS	147,001 - 151,266	27
6th highest character average	Addiction	141,740 - 157,435	27
6th lowest character average	Aortoenteric fistula	1,017 - 1,218	2
5th lowest character average	Serositis	1,125 - 1,305	2
4th lowest character average	Presbylarynx	1,191 - 1,191	0
3rd lowest character average	Omental infarction	1,142 - 1,249	3
2nd lowest character average	Saber shin	1,172 - 1,193	2
1st lowest character average	Katz syndrome	940 - 1,097	1

It is finally worth noting that, while the left panel of Fig. 6 seems to suggest that the evolution of the number of terms and references is strongly correlated, it is actually not so (Pearson's r = 0.2687). The similarity in the shape of both distributions is due to the relative low number of terms and references, and thus on the fact that some variation percentage are forbidden – to illustrate, at least 100 terms are required to have a variation between -1% and +1%.

Fig. 7 further studies the evolution of the content of snapshots, by showing how the number of the three metrics have evolved through time. Specifically, the X axis represents time (each point being an available snapshot), while the Y axis indicates the increment (or decrement) in the three metrics with respect to the initial snapshot. For the sake of clarity, and because of the high number of diseases, these are represented

as probability distributions, with each line indicating a decile (see legend in the bottom right corner). Also, for the sake of clarity, those diseases that have suffered changes in the number of characters greater than 1000 are not represented. As is to be expected, the size of the available information tends to increase over time, with only less than 10% of the diseases reducing their text.

Table 4. Number of DKEs of disease articles with higher textual content (number of characters)

Metric	Disease article	Min and max DKEs
Disease article with highest character average (1st)	Drowning	16 – 17
Disease article with highest character average (2nd)	HPV-positive oropharyngeal cancer	10 – 11
Disease article with highest character average (3rd)	Alzheimer's disease	40 – 41
Disease article with highest character average (4th)	Major depressive disorder	53 – 54
Disease article with highest character average (5th)	HIV/AIDS	28 – 30
Disease article with highest character average (6th)	Addiction	10 – 11

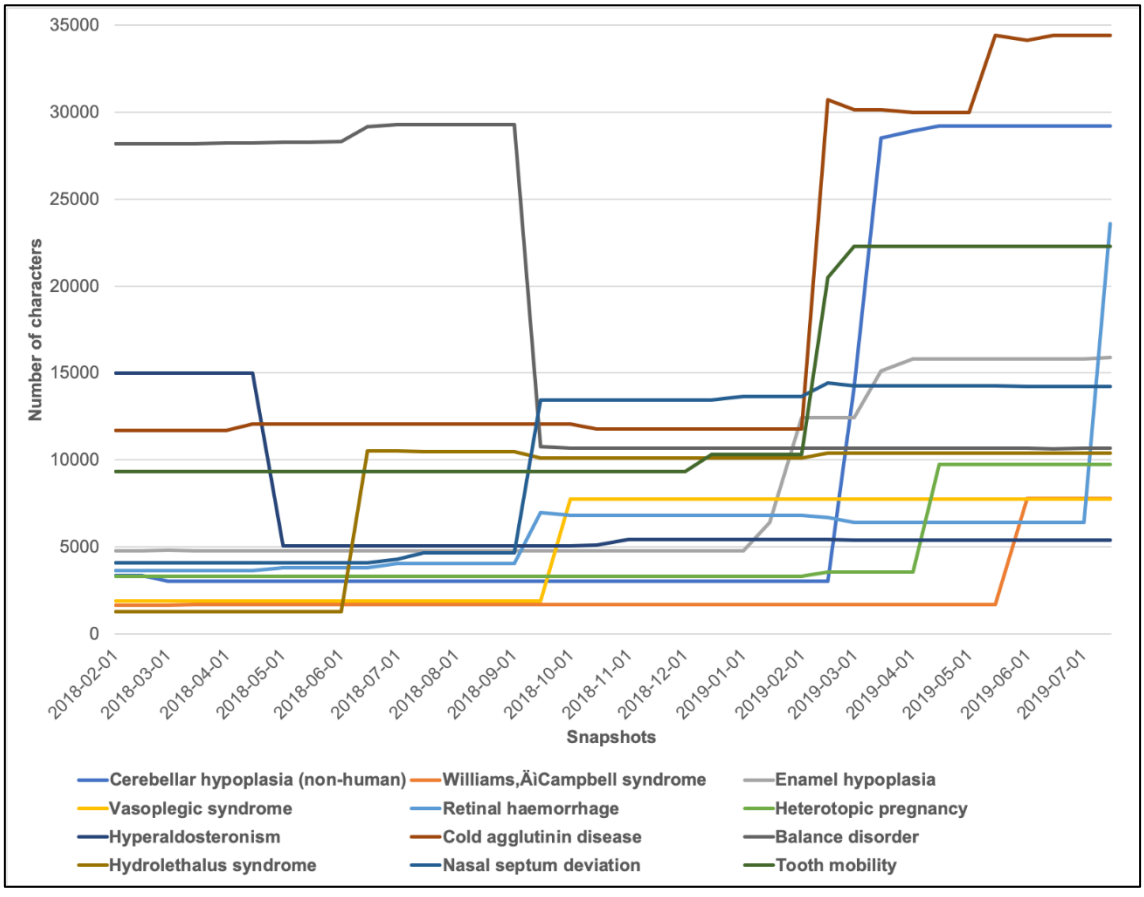

Fig. 9. Evolution through time of the number of characters for the 12 disease articles with highest variability.

We finally analyse how the three different metrics are related between them. To this aim, **Error! Reference source not found.** represents two density plots of the number of terms as a function of respectively the number of characters and of references (dark colours indicating higher densities, thus higher

number of snapshots). A clear correlation can be observed in the first case (Pearson's r = 0.5685), and a less defined one in the second (Pearson's r = 0.5093). It can therefore be concluded that, the longer the text, the higher is the number of medical terms in it included; on the other hand, terms may not always be supported by corresponding references. This may indicate that some common terms, like "fever", do not require to be supported by scientific evidence for being commonly accepted by the community.

After this global statistical analysis, we now move to the detailed study of some diseases whose snapshots present behaviours of interest.

First of all, **Table 3** presents the six diseases with the largest and smallest content, in terms of the number of characters averaged throughout all snapshots. Not surprising, in the top of the raking one can find articles like "Drowning", "Alzheimer's disease" and "HIV/AIDS", as these are of high interest both for readers and for editors: they attract large numbers of both viewers (more than 10^5 visualisations per month) and modifications. It can be appreciated that there is a strong correlation between the size of an article, and the number of times this has been edited; and that both values are probably driven by the public interest for that disease.

Along the same line, Fig. 10 presents two panels, with the evolution through time of the number of characters of respectively the top-6 and bottom-6 diseases, as reported in **Table 3**. Once again it can be appreciated that popular diseases are frequently edited, leading to a (more or less) smooth evolution. On the contrary, articles with limited content present abrupt changes, followed by long periods of inactivity. In addition to the problem of popularity, this irregular dynamic may also be due to the concentration of most modifications in a few users. Specifically, if one only considers the small articles of **Table 3**, the users registered in Wikipedia with the largest number of contributions are "Doc James"[19] a Canadian emergency physician, who edited the article referring to the "Katz syndrome" on February 13th, 2019, adding a new section and an image; and "Ozzie10aaaaaa"[20], which added, for example, two disease identifiers in the article referring to the "Aortoenteric fistula"[21] on June 2nd, 2018. Both users' history includes large number of collaborations, and recognitions by the WikiProject Medicine[22] of which they are members. In the list of contributors for the longest articles, one can also find "Doc James", along with a much more numerous lists of editors such as "Keith D"[23], "TylerDurden8823"[24] or "Boghog"[25].

Table 5. Disease articles with the highest number of DKEs.

Metric	Disease article	Min and max DKEs	Min and max characters
Disease with more DKEs (1st)	Kawasaki disease	98,0 - 98,0	62,476 – 63,193
Disease with more DKEs (2nd)	Heart failure	74,0 - 77,0	86,440 – 90,351
Disease with more DKEs (3rd)	Hypoglycemia	76,0 - 76,0	35,712 – 36,996
Disease with more DKEs (4th)	Dementia	72,0 - 75,0	124,726 – 137,119
Disease with more DKEs (5th)	Anorexia nervosa	73,0 - 75,0	77,780 – 79,216
Disease with more DKEs (6th)	Crohn's disease	75,0 - 75,0	132,236 – 137,586

The previously described trends are nevertheless not universal, as can be appreciated in **Fig. 9**, which represents the temporal evolution of the 12 diseases with more variability (in terms of character number). These articles have between 5 and 17 changes, with changes in the content size ranging from 40% to 80%. Five of them had, at some point in time, more than 20,000 characters, and can thus be considered as high-

[19] https://en.wikipedia.org/wiki/User:Doc_James
[20] https://en.wikipedia.org/wiki/User:Ozzie10aaaa
[21] https://en.wikipedia.org/w/index.php?title=Aortoenteric_fistula&oldid=844077228
[22] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine
[23] https://en.wikipedia.org/wiki/User:Keith_D
[24] https://en.wikipedia.org/wiki/User:TylerDurden8823
[25] https://en.wikipedia.org/wiki/User:Boghog

variable, large articles. Such a large variability is nevertheless not the norm: throughout the dataset there is an average variability of 10%; additionally, 3,322 disease articles show a smaller variability than the average (as opposed to 186, for which the variability is larger), indicating the presence of an asymmetrical distribution with a long right tail.

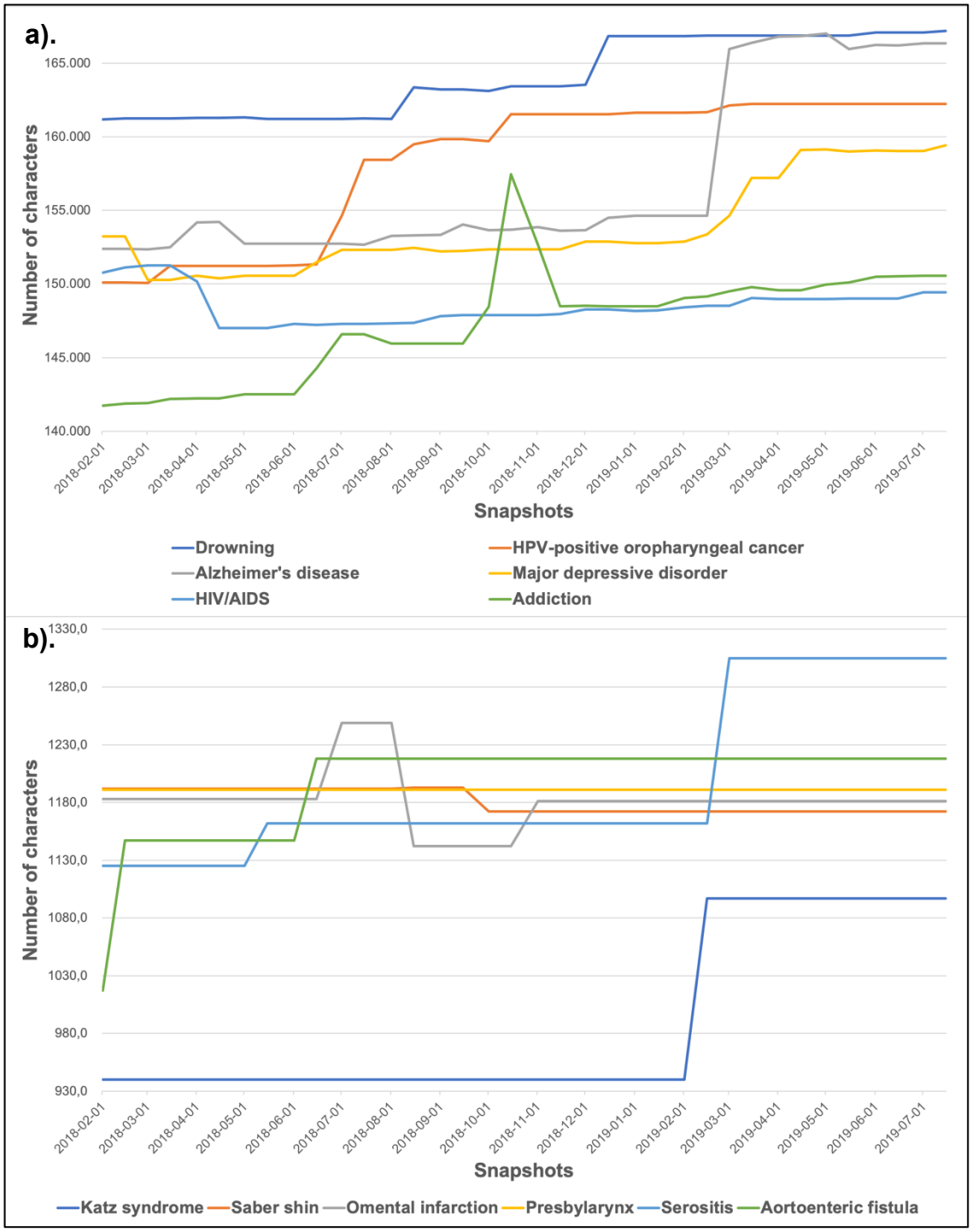

Fig. 10. Comparison between the evolution of disease articles on Wikipedia with a high and low amount of textual content.

As a last point, we analyse those articles that are most notable in terms of the number of medical terms in them contained. **Table 4** shows the number of DKEs for the six disease articles with the largest content, as per **Table 3**; and **Table 5** shows the six disease articles with the highest number of DKEs. It can be

appreciated that no strong correlation between length and number of DKEs emerges – as also previously highlighted in **Error! Reference source not found.**. This is easy to explain, as articles with the longest text may have that content distributed over different sections, not all of them relevant to diagnosis. On the other hand, short articles, as those listed in **Table 5**, may have a substantial part of them devoted to diagnosis, and may hence include a large number of terms.

5 Discussion and conclusions

This work presented an analysis of the evolution of the content of disease articles in Wikipedia, with a special focus on how such evolution may affect the automatic collection of medical knowledge and its use in subsequent data science tasks. Consequently, three metrics were considered, i.e. overall text length, number of medical terms (DKEs), and number of references.

Some general conclusions can be drawn, all of them consistent with the intuition the average user of Wikipedia may have. Knowledge globally increases with time, both in terms of characters, terms and references; only 10% of diseases have seen an overall reduction in their content. The size of articles also correlates with the popularity of the corresponding diseases, such that “hot topics” are both visualised and edited more frequently. Finally, the longer the text, the greater the number of DKEs, even though the latter ones might not always be supported by the corresponding references.

The results here presented also highlight the importance of the temporal dimension in any analysis based on Wikipedia, and how a special care should be devoted to “minor” articles. Small articles are seldom updated, but when this happens, important changes in the content may be introduced. It is therefore important to use recent datasets, as the amount of information may double in a matter of days. Small articles are also small because they refer to niche topics, and hence receive the attention of a restricted number of editors and viewers. This implies a higher risk of biased or wrong information, as less editors collaborate on the writing process, and less readers may be able to raise a flag.

If small articles are potentially less reliable, long ones also presents drawbacks. While their behaviour is usually more stable, they can still suffer from important abrupt changes. An interesting example is provided by “Cerebellar hypoplasia”, for which the content has halved between two consecutive snapshots – see Fig. 10. Additionally, long texts in the article do not imply the presence of a large number of medical terms, nor of references, which may limit their usefulness in automatic data extraction tasks.

Some limitations of this work have to finally be discussed. First of all, not all articles identified in Wikipedia as diseases are such, as for instance "Subprime mortgage crisis" or "Domestic violence". Still, it is worth noting that there is no universally accepted definition of diseases in bioinformatics, but only a large number of (not always consistent) taxonomies. Additionally, some diseases do not apply to humans, as is the case of "Cerebellar hypoplasia (non-human)". Beyond the previously listed limitations of Wikipedia, the prospective user should consider the use of filters, based on external data sources, to avoid such cases. Secondly, while only textual information has here been considered, Wikipedia’s articles include information in other medias, which may also be of relevance. To illustrate, one can consider the case of “Madelung's deformity”. Its article only includes one DKE (i.e., pain); yet, it is accompanied by an image that clearly explains the deformity, up to the point that even a person without medical background could diagnose it. A complete system for knowledge extraction should thus also consider non-textual information, whenever possible.

Funding

The paper is a result of the project “DISNET (Creation and analysis of disease networks for drug repurposing from heterogeneous data sources applied to rare diseases)”, that is being developed under grant “RTI2018-094576-A-I00” from the Spanish Ministerio de Ciencia, Innovación y Universidades. Gerardo Lagunes-Garcia work is supported by Mexican Consejo Nacional de Ciencia y Tecnología (CONACYT) (CVU: 340523) under the programme "291114 - BECAS CONACYT AL EXTRANJERO".

References

Al Tamime, R. (2017). Behind The Wikipedia Medical Knowledge Factory: Understanding the Knowledge Dynamic Over Time. Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 238–243. https://doi.org/10.1145/3027063.3027119

Allahwala, U. K., Nadkarni, A., & Sebaratnam, D. F. (2013). Wikipedia use amongst medical students – New insights into the digital revolution. Medical Teacher, 35(4), 337–337. https://doi.org/10.3109/0142159X.2012.737064

Azer, S. A. (2015). Is Wikipedia a reliable learning resource for medical students? Evaluating respiratory topics. Advances in Physiology Education, 39(1), 5–14. https://doi.org/10.1152/advan.00110.2014

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32(suppl_1), D267–D270. https://doi.org/10.1093/nar/gkh061

Boudoulas, K. D., Triposkiadis, F., Stefanadis, C., & Boudoulas, H. (2017). The endlessness evolution of medicine, continuous increase in life expectancy and constant role of the physician. Hellenic Journal of Cardiology, 58(5), 322–330. https://doi.org/10.1016/j.hjc.2017.05.001

Cohen, N. (2013, September 29). Editing Wikipedia Pages for Med School Credit. The New York Times. Retrieved from https://www.nytimes.com/2013/09/30/business/media/editing-wikipedia-pages-for-med-school-credit.html

Densen, P. (2011). Challenges and Opportunities Facing Medical Education. Transactions of the American Clinical and Climatological Association, 122, 48–58.

Espe, S. (2018). Malacards: The Human Disease Database. Journal of the Medical Library Association : JMLA, 106(1), 140–141. https://doi.org/10.5195/jmla.2018.253

Fairchild, G., Del Valle, S. Y., De Silva, L., & Segre, A. M. (2015). Eliciting Disease Data from Wikipedia Articles. Proceedings of the ... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media, 2015, 26–33.

Fernandez-Moure, J. S. (2016). Lost in Translation: The Gap in Scientific Advancements and Clinical Application. Frontiers in Bioengineering and Biotechnology, 4. https://doi.org/10.3389/fbioe.2016.00043

Friedlin, J., & McDonald, C. J. (2010). An evaluation of medical knowledge contained in Wikipedia and its use in the LOINC database. Journal of the American Medical Informatics Association: JAMIA, 17(3), 283–287. https://doi.org/10.1136/jamia.2009.001180

García, G. Lagunes, Santamaria, L. P., Valle, E. P. G. del, Zanin, M., Ruiz, E. M., & González, A. R. (2019). Wikipedia Disease Articles: An Analysis of their Content and Evolution. 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), 664–671. https://doi.org/10.1109/CBMS.2019.00136

García, Gerardo Lagunes, González, A. R., Santamaría, L. P., Valle, E. P. G. del, Zanin, M., & Ruiz, E. M. (2018). DISNET: Disease understanding through complex networks creation and analysis. BioRxiv, 428201. https://doi.org/10.1101/428201

Goslin, K., & Hofmann, M. (2018). A Wikipedia powered state-based approach to automatic search query enhancement. Information Processing & Management, 54(4), 726–739. https://doi.org/10.1016/j.ipm.2017.10.001

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research, 33(Database Issue), D514–D517. https://doi.org/10.1093/nar/gki033

Hodson, R. (2015). Wikipedians reach out to academics. Nature News. https://doi.org/10.1038/nature.2015.18313

Jiang, Y., Bai, W., Zhang, X., & Hu, J. (2017). Wikipedia-based information content and semantic similarity computation. Information Processing & Management, 53(1), 248–265. https://doi.org/10.1016/j.ipm.2016.09.001

Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., … Robinson, P. N. (2017). The Human Phenotype Ontology in 2017. Nucleic Acids Research, 45(D1), D865–D876. https://doi.org/10.1093/nar/gkw1039

Lagunes García, G., Prieto Santamaria, L., Garcia del Valle, E. P., Zanin, M., Menasalvas Ruiz, E., & Rodríguez González, A. (2019). Wikipedia Disease Articles: An Analysis of their Content and Evolution. 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), 664–671. https://doi.org/10.1109/CBMS.2019.00136

Lindberg, D. A. (2000). Internet access to the National Library of Medicine. Effective Clinical Practice: ECP, 3(5), 256–260.

Lipscomb, C. E. (2000). Medical Subject Headings (MeSH). Bulletin of the Medical Library Association, 88(3), 265–266.

Mayo Clinic. (2019). Retrieved February 16, 2018, from https://www.mayoclinic.org/

Mehdi, M., Okoli, C., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2017). Excavating the mother lode of human-generated text: A systematic review of research that uses the wikipedia corpus. Information Processing & Management, 53(2), 505–529. https://doi.org/10.1016/j.ipm.2016.07.003

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., & Lanamäki, A. (2015). "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. Journal of the Association for Information Science and Technology, 66(2), 219–245. https://doi.org/10.1002/asi.23172

Miller, N., Lacroix, E.-M., & Backus, J. E. B. (2000). MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service. Bulletin of the Medical Library Association, 88(1), 11–17.

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., … Furlong, L. I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research, 45(D1), D833–D839. https://doi.org/10.1093/nar/gkw943

Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X., & Jensen, L. J. (2015). DISEASES: Text mining and data integration of disease–gene associations. Methods, 74, 83–89. https://doi.org/10.1016/j.ymeth.2014.11.020

Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Iny Stein, T., … Lancet, D. (2013). MalaCards: an integrated compendium for diseases and their annotation. Database, 2013. https://doi.org/10.1093/database/bat018

Rappaport, N., Twik, M., Nativ, N., Stelzer, G., Bahir, I., Stein, T. I., … Lancet, D. (2014). MalaCards: A Comprehensive Automatically-Mined Database of Human Diseases. Current Protocols in Bioinformatics, 47(1), 1.24.1-1.24.19. https://doi.org/10.1002/0471250953.bi0124s47

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. American Journal of Human Genetics, 83(5), 610–615. https://doi.org/10.1016/j.ajhg.2008.09.017

Safran, M., Solomon, I., Shmueli, O., Lapidot, M., Shen-Orr, S., Adato, A., … Lancet, D. (2002). GeneCardsTM 2002: towards a complete, object-oriented, human gene compendium. Bioinformatics, 18(11), 1542–1543. https://doi.org/10.1093/bioinformatics/18.11.1542

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., … Kibbe, W. A. (2012). Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Research, 40(D1), D940–D946. https://doi.org/10.1093/nar/gkr972

Shafee, T., Masukume, G., Kipersztok, L., Das, D., Häggström, M., & Heilman, J. (2017). Evolution of Wikipedia's medical content: past, present and future. J Epidemiol Community Health, 71(11), 1122–1129. https://doi.org/10.1136/jech-2016-208601

Thompson, N., & Hanley, D. (2017). Science Is Shaped by Wikipedia: Evidence from a Randomized Control Trial (SSRN Scholarly Paper No. ID 3039505). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=3039505

Tucker, M. E. (2014, February 5). Doctors, Not Just Patients, Use Wikipedia, Too: IMS Report. Retrieved September 29, 2019, from Medscape website: http://www.medscape.com/viewarticle/820249

Valle, E. P. G. del, García, G. L., Santamaría, L. P., Zanin, M., Ruiz, E. M., & González, A. R. (2018). Evaluating Wikipedia as a Source of Information for Disease Understanding. 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), 399–404. https://doi.org/10.1109/CBMS.2018.00076

Wang, J. Z., Pourang, A., & Burrall, B. (2019). Open access medical journals: Benefits and challenges. Clinics in Dermatology, 37(1), 52–55. https://doi.org/10.1016/j.clindermatol.2018.09.010