Wikipedia disease articles: an analysis of their content and evolution

Gerardo Lagunes García
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Pozuelo de Alarcon, Madrid, Spain
gerardo.lagunes@ctb.upm.es

Lucia Prieto Santamaria
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Pozuelo de Alarcón, Madrid, Spain
lucia.prieto.santamaria@alumnos.upm.es

Eduardo P. Garcia del Valle
ETS Ingenieros Informáticos
Universidad Politécnica de Madrid
Pozuelo de Alarcon, Madrid, Spain
ep.garcia@alumnos.upm.es

Massimiliano Zanin
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Pozuelo de Alarcón, Madrid, Spain
massimiliano.zanin@upm.es

Ernestina Menasalvas Ruiz
Centro de Tecnología Biomédica, ETS Ingenieros Informáticos
Universidad Politécnica de Madrid
Madrid, Spain
ernestina.menasalvas@upm.es

Alejandro Rodríguez González
Centro de Tecnología Biomédica, ETS Ingenieros Informáticos
Universidad Politécnica de Madrid
Madrid, Spain
alejandro.rg@upm.es

***Abstract*— Nowadays there is a huge amount of medical information that can be retrieved from different sources, both structured and unstructured. Internet has plenty of textual sources with medical knowledge (books, scientific papers, specialized web pages, etc.), but not all of them are publicly available. Wikipedia is a free, open and worldwide accessible source of knowledge. It contains more than 150,000 articles of medical content in the form of texts (non-structured information) that can be mined. The aim of this work is to study whether the information contained in Wikipedia medical articles can be used in a research context. The study has been focused on extracting the elements, from Wikipedia disease articles, that can be used to guide a diagnosis process, support the creation of diagnostic systems, or analyze the similarities between diseases, among others. The results provided show that Wikipedia is a rich source of diagnostic knowledge that can be exploited and used in research.**

Keywords— wikipedia; diseases; diagnosis; wikipedia evolution

I. INTRODUCTION

Wikipedia is an online source of information, open and collaborative. At the moment of analysis, it contains more than 43 million of pages, more than 5 million articles (in English)[1] and is released in 298 languages (10 not used)[2]. It is one of the most visited web sites[3], being the English version the biggest and most active.

In March 2017, there were 30,000 medical articles in the English Wikipedia [1], which increased to 36,850 articles in March 2018 [2], being one of the most used medical sources by general population and by specialists in medicine [1] that have deeply collaborated for their enrichment [3]. In 2014 Wikipedia was referred as "the single leading source of medical information for patients and health care professionals" by the Institute of Medical Science (IMS) Institute for Healthcare Informatics, as stated in [4]. Wikipedia is a public and collaborative encyclopaedia which encourages the inclusion and the edition of the content published by constant updates. For this reason, the content of Wikipedia is adapted to the day-by-day reality, a noteworthy feature that has raised the interest of the scientific community.

As has been stated, Wikipedia has a considerable number of published medical articles, including information about diseases. The information about diseases normally contains one or more sections that provide information about what are the main medical elements (signs and symptoms, diagnostic tests, laboratory test results, etc.) that are used to guide the diagnosis of a specific disease. These elements, which in the case of the symptoms and signs can be seen as the phenotypic manifestation of a disease, can be of interest in several research fields, and are especially relevant in tasks like the creation of

[1] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia
[2] https://en.wikipedia.org/wiki/List_of_Wikipedias
[3] https://www.alexa.com/topsites

disease-phenotype knowledge bases [5] or of automated diagnostic systems [6] [6]. In spite of this, little attention has been devoted to the use of Wikipedia as a source of diagnostic information, and to its validation.

This paper tackles this issue, by analysing the diagnostic information contained in Wikipedia medical articles, and by showing the results of applying a text-mining process over Wikipedia disease articles. The paper studies the relevance of Wikipedia evaluating several metrics such as: Wikipedia updates, number of disease articles and findings, information retrieved, and related.

II. Related Work

There are several research works in the biomedical field that have used Wikipedia as their main information source. In 2013 a paper was published [7], describing a web application created to store information about diseases and their symptoms as extracted from Wikipedia, with the aim of detecting diseases suffered by specific users based on the tracking of the searches performed on the Web. With data from the same year (2013-2014), in 2015 a second study was published [8] that used Wikipedia to forecast the influenza outbreak. For that, the authors did a study of the registry access to Wikipedia articles. In the same line, in 2014 was created a system [9] for the monitoring and prognosis of diseases by, again, analysing the access to Wikipedia articles. In 2015 a study made use of the "2014 West African Ebola virus disease epidemic" article to get information regarding death counts and hospitalization counts in the narratives; it further proposed the use of Wikipedia as a community-driven open-source emerging disease, detection, monitoring and repository system, with the rationale that current surveillance systems suffer from disadvantages such as reporting lags and antiquated technology.

Other works have been focused on the improvement of Wikipedia articles by means of Natural Language Processing (NLP) techniques, and by applying automatic evaluation of Wikipedia medical articles [10].

In the diagnosis context, a system was proposed to automatically infer the most probable diagnosis from clinical narratives [11]. The authors tested their system with texts from Wikipedia, Mayo clinic, Freebase and UMLS, and found that Wikipedia and Mayo clinic-based systems reached respectively a 60% and 70% of correct diagnoses, thus suggesting that those sources were very relevant for finding correct diagnosis. In a similar approach, very aligned with our study, a recent work has studied the feasibility of using Wikipedia for extracting disease terms, aimed at disease understanding [12], with promising results.

Other studies have been focused in the creation of medical ontologies, using repositories such as Wikipedia, under the assumption that the latter "*provides a valuable resource from which to mine structured information*" [13]. Another study in the same direction used Wikipedia for the creation of a clinical thesaurus [14].

The knowledge contained in Wikipedia has been also used to enrich SNOMED-CT[4] to obtain medical terms synonyms. As a result of this approach, 183,100 new synonyms were retrieved with an accuracy of 85.6%, demonstrating again the powerful value of the knowledge contained in Wikipedia [15].

The analysis of the related work leads to the conclusion that Wikipedia has been used as a trustful source of knowledge. It contains online information that can be retrieved and used for many purposes, including medical research ones. Using Wikipedia to obtain medical terms regarding the diagnosis of diseases is an unusual, different, interesting and unique approach.

III. Materials and Methods

The aim of the study is to perform an analysis of the "diagnostic knowledge" information contained in Wikipedia disease articles. For this work, we consider as diagnostic knowledge all the elements that are related to a disease and allow physicians to guide the diagnosis process. Stricto sensu, these include the phenotypic manifestations, i.e. findings, signs and symptoms. However, other elements such as diagnostic procedures, laboratory tests and results are also considered diagnostic knowledge, as they allow guiding the diagnosis process.

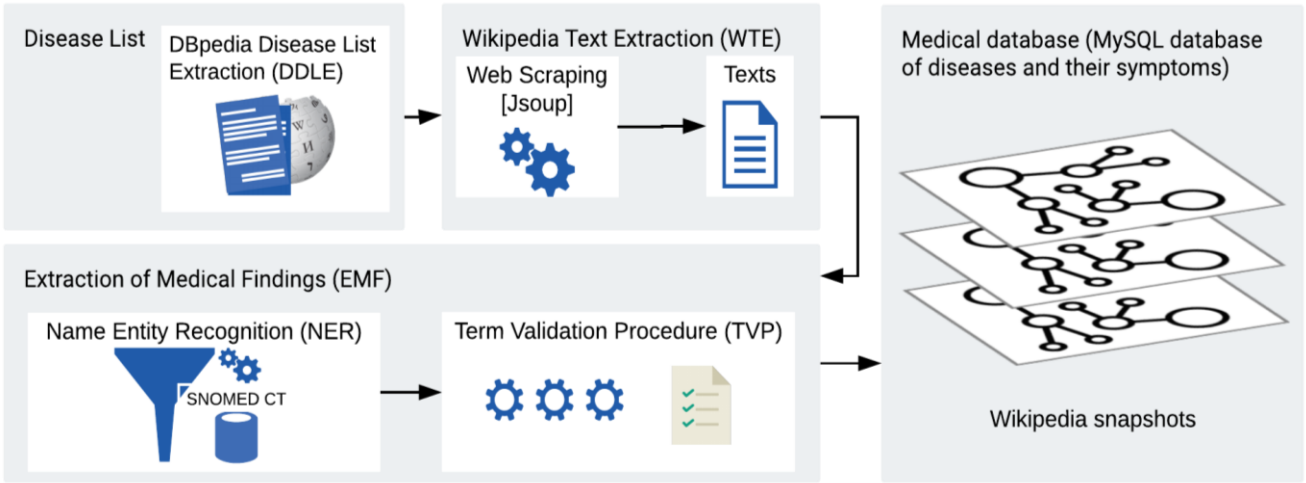

Fig. 1. Pipeline extraction process

We have created a pipeline (Figure 1) that is executed two times per month to extract the diagnostic knowledge contained in Wikipedia disease articles. This process allows the creation of snapshots of the Wikipedia data, thus enabling the study of their evolution. The pipeline performs the following steps: i) retrieves a list of available Wikipedia diseases articles; ii) retrieves texts from Wikipedia articles; iii) performs text-mining over texts and iv) analyses the obtained results.

A. Information Source

The information source is composed of all Wikipedia articles categorized as diseases. The diseases list (and disease codes) were extracted from DBPedia [16] using a SPARQL query. A disease Wikipedia article is, in most of the cases, structured using different sections that includes description, disease codification, causes, diagnosis, treatment, prognosis and clinical manifestations. For extracting diagnostic knowledge, the relevant sections are: "*Signs and symptoms", "Causes", "Diagnosis"* and *"Presentation"*. The information retrieved was: 1) the texts contained in the aforementioned sections; 2) links contained in the texts and 3) disease codifications.

B. Data retrieval and knowledge extraction

The pipeline for extracting information has been executed twice per month, being the first execution on February 1st, 2018 and the last one reflected in this study on August 15th, 2018, thus yielding a total of 14 snapshots. The list of Wikipedia disease articles was retrieved executing the DBpedia Disease List Extraction (DDLE) process: a procedure that performs a SPARQL query[5] against DBpedia. The second step consists in a process named Wikipedia Text Extraction (WTE) developed using Jsoup API[6] (see [17] for further information), which extracts the texts from the sections using web-scrapping and stores them in a MySQL database.

The next step applies the Extraction of Medical Findings (EMF) process: a NLP procedure to retrieve the relevant elements. This process is based in a two-step approach. Firstly, MetaMap [18] performs a Name Entity Recognition (NER) over the texts and retrieves the relevant ones based on the arguments provided (sources to be used to identify terms and UMLS semantic types[7] to be detected). Secondly, our Validation Procedure (TVP) module, described in [17], is applied. TVP validates the terms retrieved by MetaMap minimizing the number of terms incorrectly detected. Finally, all results are stores in a MySQL database.

The source used to identify medical terms in the NER process was SNOMED_CT in English. The semantic types passed as argument to MetaMap were: *sosy*, *diap*, *dsyn*, *fndg*, *lbpr*, *lbtr*, *inpr*, *menp*, *mobd*, *patf* and *cgab*. Regarding semantic types, from now on all the elements found by the EMF process will be named as Diagnostic Knowledge Element (DKE), independently on the associated semantic type.

IV. Results

The execution of the pipeline yielded 14 snapshots from February 1st, 2018 until August 15th, 2018. Table I reports some key properties describing the last snapshot; similar information for the other time points is available in Table V.

TABLE I. VALUES OF KEY PARAMETER THAT MAKE UP THE AUGUST 15TH, 2018 SNAPSHOT

Key properties	Count
Articles categorized as disease in DBpedia (DDLE process)	9,858
Articles that contained diagnostic knowledge elements (from EMF process)	4,372
Diagnostic knowledge elements (DKE) found (no duplicates)	12,748
UMLS semantic types found (no duplicates)	15
Disease codes found (no duplicates)	19,226
External vocabularies found (no duplicates)	60
Number of texts (no duplicates)	35,525
Number of links found in the texts	168,404

As can be observed there is a significant difference between the number of articles categorized in DBpedia as diseases (9,858) and the number of those whose content could be retrieved (4,372 – 44.34%). There are two explanations for this phenomenon: i) the articles did not contain diagnostic knowledge elements; or ii) the articles are incorrectly catalogued – i.e. in DBpedia these were catalogued with the class "Disease", but Wikipedia real article did not contain information about a disease. To find out whether these differences came from, we have assumed that any article about a disease should have, at least, one code of an external vocabulary/classification system (ICD, OMIM, etc.) in the data retrieved from DBpedia (DDLE process) or in the data retrieved from the scrapping of Wikipedia (WTE process). Those articles that return 0 codes in both sets have been classified as non-disease articles, implying that DBpedia has catalogued them incorrectly. More information about those articles that have been retrieved as diseases in DBpedia but have not been considered as diseases is available online[8].

The number of external vocabularies/classification systems found in Wikipedia articles is 71[9]. This information is relevant in order to get more information about a specific disease in an

[5] https://github.com/GerardoUPM/wikipediaAnalysisMedicalData/blob/master/getDiseases.sparql

[6] https://jsoup.org/

[7] https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

[8] https://github.com/GerardoUPM/wikipediaAnalysisMedicalData/tree/master/articles_that_have_been_retrieved_as_diseases_in_DBpedia_but_have_not_been_considered_as_diseases

[9] https://github.com/GerardoUPM/wikipediaAnalysisMedicalData/blob/master/External_Vocabulary_Found.csv

external database. Results indicate that Wikipedia is not only useful because of its content, but also as a bridge to reach external sources.

TABLE II. DISEASE ARTICLES WITH MORE AND LESS DISEASE DKES (SNAPSHOT: AUGUST 15TH, 2018)

Metric	Disease	URL http://en.wikipedia.org	Number of unique DKEs
Disease with more DKE (1st)	Kawasaky disease	/wiki/Kawasaki_disease	98
Disease with more DKE (2nd)	Hypoglycemia	/wiki/Hypoglycemia	75
Disease with more DKE (3rd)	Anorexia nervosa	/wiki/Anorexia_nervosa	74
Disease with less DKE (1st)	ABCD syndrome	/wiki/ABCD_syndrome	1
Disease with less DKE (2nd)	Abietic acid dermatitis	/wiki/Abietic_acid_dermatitis	1
Disease with less DKE (3rd)	Accessory breast	/wiki/Accessory_breast	1

With respect to the scope of this work, the main metric is the number of retrieved diagnostic knowledge elements, which totalled 12,748. Table II reports additional information about this metric, including the three diseases with more and less DKE. The average number of DKE is 11.79.

Regarding the specific DKEs that have been found, Table III shows the more and less frequent terms.

TABLE III. MORE AND LESS FREQUENT DKES (TOTAL APPEARANCES OF THE TERM; TOTAL DISEASE ARTICLES WHERE THE TERM APPEARS) (SNAPSHOT: AUGUST 15HT, 2018)

Metric	Term	Number of appearances
Term with more appearances (1st)	Pain (C0030193)	1,676
Term with more appearances (2nd)	Lesion (C0221198)	1,157
Term with more appearances (3rd)	Magnetic resonance imaging (C0024485)	866
Term with less appearances (1st)	Catatonic stupor (C0233607)	1
Term with less appearances (2nd)	Benign essential blepharospasm (C2930898)	1
Term with less appearances (3rd)	Numbness of finger (C0587054)	1

Another measured metric was the number of UMLS semantic types found. These semantic types correspond to a division of DKEs into homogeneous and conceptually-related families. The semantic types were filtered to 11 in the EMF process, but the process retrieved 17: Acquired Abnormality (*acab*), Anatomical Abnormality (*anab*), Congenital Abnormality (*cgab*), Diagnostic Procedure (*diap*), Disease or Syndrome (*dsyn*), Finding (*fndg*), Intellectual Product (*inpr*), Laboratory Procedure (*lbpr*), Laboratory or Test Result (*lbtr*), Molecular Biology Research Technique (*mbrt*), Mental Process (*menp*), Manufactured Object (*mnob*), Mental or Behavioral Dysfunction (*mobd*), Pathologic Function (patf), Quantitative Concept (*qnco*), Tissue (*tisu*) and Sign or Symptom (*sosy*). The reason behind the higher number of semantic types retrieved by EMF, as opposed to those imposed as parameters, is the fact that a term can belong to more than one semantic type. The list of the 12,748 distinct DKEs that have been retrieved in all the snapshots with their corresponding semantic types is also available online[10]. The distribution of the DKEs grouped by their semantic type is available in Table IV.

TABLE IV. DISTRIBUTION OF DKES GROUPED BY THEIR SEMANTIC TYPE (SNAPSHOT: AUGUST 15TH, 2018)

Semantic type	Count
acab	13
anab	35
cgab	850
diap	554
dsyn	5,604
fndg	2,623
inpr	304
lbpr	674
lbtr	173
mbrt	1
menp	165
mnob	2
mobd	537
patf	775
sosy	804

As can be seen, the semantic types with the highest number of DKE are *fndg*, *dsyn*, *cgab* and *sosy*, something that was to be expected given the type of information that we wanted to retrieve. On the other side, we have *mbrt* and *mnob* as the less frequent semantic types.

[10] https://github.com/GerardoUPM/wikipediaAnalysisMedicalData/blob/master/DKE_RetrievedInTheSnapshotOf_2018-08-15.csv

TABLE V. EVOLUTION OF THE INFORMATION CONTAINED IN THE SNAPSHOTS

Snapshot	DBpDis	WRDArt	WRawDF	WTxt	WST	WExCd	WExtSrc	WLink
2018-02-01	8,161	3,881	9,937	31,126	11	19,229	64	149,368
2018-02-15	8,161	3,889	9,953	31,203	11	19,187	63	149,802
2018-03-01	8,161	3,907	11,697	31,393	15	19,141	62	150,528
2018-03-15	8,161	3,917	11,731	31,657	15	19,127	63	151,794
2018-04-01	9,857	4,155	11,889	33,203	15	19,327	65	158,574
2018-04-15	9,858	4,162	12,408	33,263	16	19,314	65	159,097
2018-05-01	9,858	4,177	12,432	33,400	15	19,297	65	159,900
2018-05-15	9,858	4,187	12,479	33,490	15	19,295	65	160,201
2018-06-01	9,858	4,332	12,615	34,628	15	19,271	61	164,727
2018-06-15	9,858	4,341	12,642	34,836	15	19,254	62	165,726
2018-07-01	9,858	4,345	12,709	35,218	15	19,239	63	166,764
2018-07-15	9,858	4,352	12,722	35,321	15	19,237	62	167,221
2018-08-01	9,858	4,366	12,743	35,430	15	19,232	62	167,909
2018-08-15	9,858	4,372	12,748	35,525	15	19,226	60	168,404

Regarding information evolution, we have measured several metrics in the different snapshots: number of articles retrieved by DBpedia as diseases (DBpDis), number of Wikipedia articles that contain DKE (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). Table V shows the evolution of the metrics in the different snapshots.

V. DISCUSSION

Before starting a discussion about the obtained results, it is necessary to mention how this information has been validated. First, we have relied on the accuracy of MetaMap as NER tool. MetaMap has been widely tested, and is considered as a valid system to perform NLP processes, and more specifically NER over medical texts [19] [20] [21] [22]. On the other hand, the data pipeline includes the validation of the terms retrieved by MetaMap using our TVP module [17]. This latter element has been previously validated using MedLine Plus medical texts as information source, which are content and structure-wise qualitatively similar to Wikipedia's ones. Due to this, it was estimated that an ex-novo validation was not necessary.

Time plays an important role because it allows to observe the evolution of the knowledge stored in Wikipedia. As can be observed in Table V, there is a progression in most of the elements. Only Semantic Types (WST) remains with no significant changes – as is to be expected, since this number could only change by finding new DKEs of different types. All metrics (except the WExCd, which presents an irregular behavior) are monotonically increasing, thus suggesting that the modifications of the Wikipedia articles result in the inclusion of new information that is captured by our pipeline. Fig. 2 shows a graphical representation of the number of Wikipedia articles that contain DKE (WRDArt), the number of DKE found by MetaMap (not applying TVP validation but removing duplicates) (WRawDF), number of texts (WTxt) and number of external codes found in Wikipedia (WExCd). In each graph we see a trend line of evolution.

From a global point of view, all these metrics support the feasibility of using Wikipedia as a source of medical information. We have been able to obtain information from 4,440 articles that are catalogued as diseases and contained diagnostic information, of which 4,163 were diseases with at least one medical term (after applying TVP). While noteworthy, this number should be compared with those obtained in other medical sources: ≈4,500 for MeSH[11][23]; ≈8,500 for OMIM[12][24]; and ≈10,500 for DisGeNET[13]. While prima facie DisGeNET includes information about twice the diseases of Wikipedia, two additional aspects have to be taken into account. First of all, Wikipedia is an alive system, frequently updated by its users, and whose knowledge base is constantly increasing - as shown in Table V. Secondly, different websites rely on different vocabularies: a same disease can then be classified differently, split into different subtypes, merged, and so forth. Therefore, the high number

[11] https://github.com/GerardoUPM/wikipediaAnalysisMedicalData/blob/master/getDiseasesMesh.sparql

[12] https://www.omim.org/statistics/entry

[13] https://github.com/GerardoUPM/wikipediaAnalysisMedicalData/blob/master/getDiseasesDisGeNET.sparql

found for DisGeNET does not necessarily imply a larger body of information.

The average number of obtained disease findings is another important factor. 1,709 (38.49%) diseases have at least 11 identified disease findings. This is positive from the point of view of creating, for instance, diagnosis systems based on the information here retrieved. Online we can look at the list of these diseases and their number of disease findings[14].

In Fig. 3 we see a graph comparing the first generated snapshot against the last one. We confirm that during these last seven months we have found an increase in the information contained in the articles catalogued as diseases and, above all, that each new contribution has enough medical textual content to allow us to identify an increasing number of disease findings.

In spite of all the above, there are a few issues that are worth discussing. First of all, we have found many articles that were catalogued as diseases in DBpedia, but that were discarded after applying our filter method (no external codes in DBpedia nor Wikipedia). Some of them are related to the medical domain (for example: "famous outbreaks", "health crisis", etc.), but they are not "diseases". Some examples of those articles are "1924 Los Angeles pneumonic plague outbreak", "1852–60 cholera pandemic"; "1863–75 cholera pandemic" and "2013 Swansea measles epidemic". However, many others are not even related with medical information in any way (for example: "2008 Western Australian gas crisis", "2010 in film", "2003 Wimbledon Championships – Women's Singles"). Another drawback is the detection of potential relevant Wikipedia articles that have returned no results in the EMF process. The reason lies in the fact that we have limited the text-mined sections to "*Signs and symptoms", "Causes", "Diagnosis" and "Presentation"*, but these are empty in some articles (e.g.: "Hereditary sensory and autonomic neuropathy", "17q21.31 microdeletion syndrome", "2,4 Dienoyl-CoA reductase deficiency"). On the other hand, we have found disease articles that have not been structured using the sections that we are using to retrieve terms (e.g.: "Bleb (medicine)", "Oligodactyly" o "Meteoropathy"). Finally, one should take care of articles referred to diseases that are no longer catalogued as such (e.g.: "Female Hysteria").

However, even taking those drawbacks into account, results are promising: if we discard the Wikipedia articles not considered as disease based on our filter (9,858) we have found medical findings in 4,372 (44.34%), which represents a really high number. The numbers that have here been presented and analysed also provide a good benchmark for the aim established: obtain as much diagnostic information as possible from Wikipedia articles.

Finally, we can conclude that Wikipedia is an accurate and relevant source of medical information due to the collaborative quality control that Wikipedia nowadays enforces, the introduction of wrong information is really difficult, and the creation of snapshots and the analysis of the difference between them could allow us to discard irrelevant data. On the other hand, the constant updates make Wikipedia a highly up to date source of information – an essential feature for research.

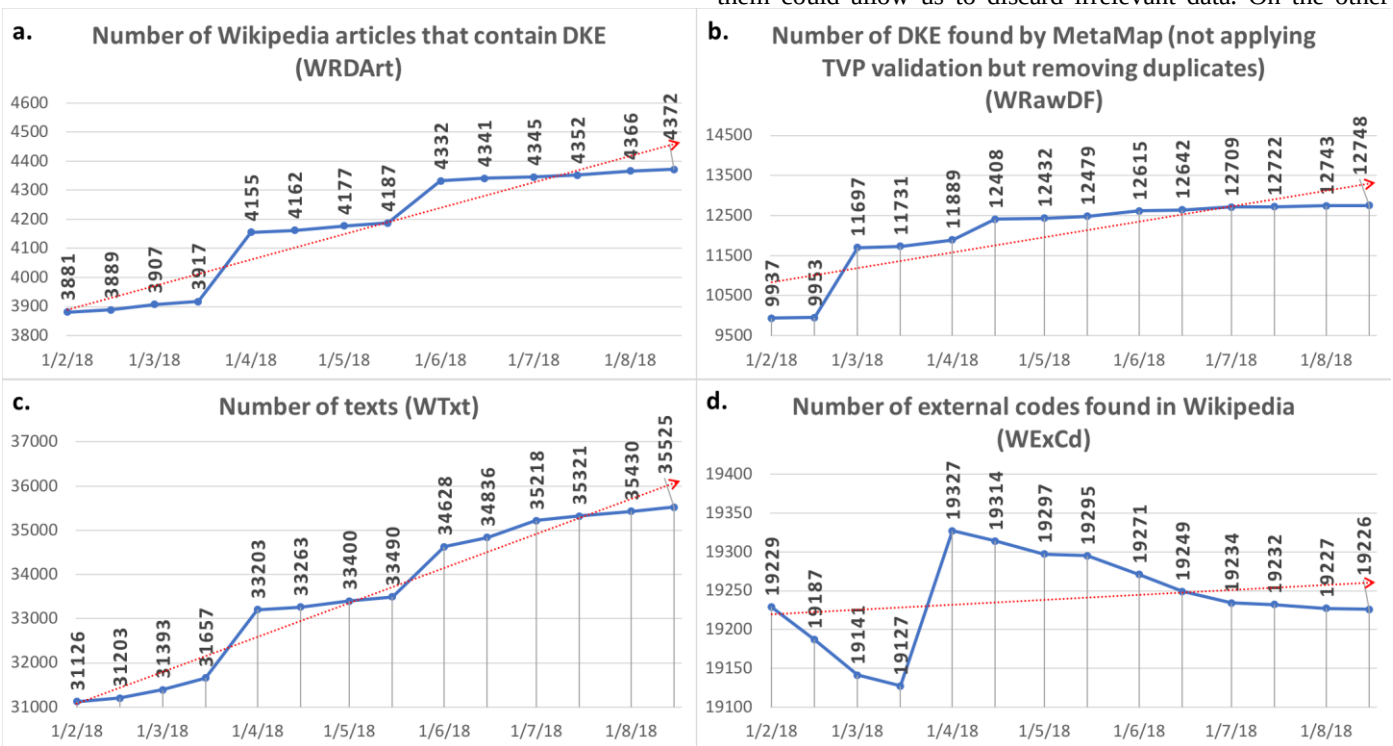

Fig. 2 Graphical representation of the key parameters WRDArt, WRawDF, WTxt and WExCd

[14] https://github.com/GerardoUPM/wikipediaAnalysisMedicalData/blob/master/Diseases_and_their_number_of_disease_findings.csv

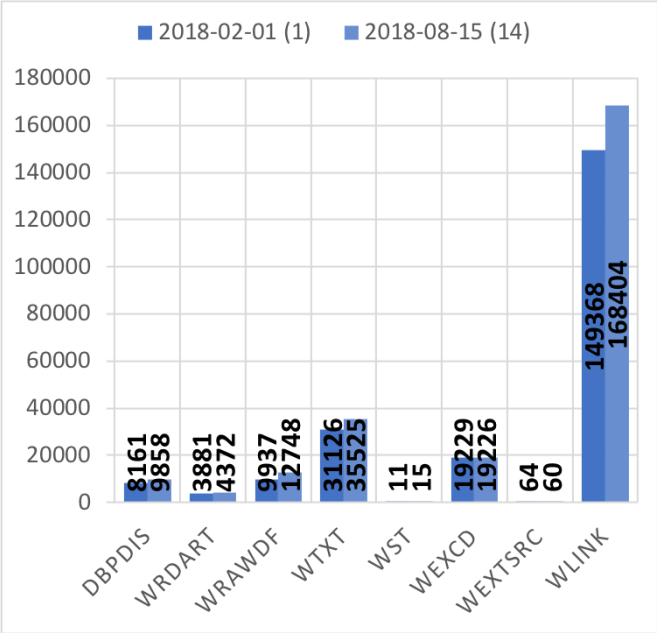

Fig. 3 Comparison between the first snapshot (February 1st, 2018) and the last snapshot (August 15th, 2018)

VI. Future Work

Future work will be focused on a three-fold strategy.

First of all, we consider that the information extracted to perform this study can be a valuable resource for researchers working in fields like disease understanding, disease networks or diagnosis systems. For this reason, we are currently developing a platform that will allow accessing all the data that have been processed and summarized in this study. Secondly, effort will be devoted to solve the drawbacks that were here found, including: improving and applying the filter method to discard no-disease articles; improving the extraction of texts for those articles without the predefined sections; and discarding deprecated articles.

A final future work will be to perform this analysis again in a longer time scale, trying to analyse the same data with more information, and trying also to find an explanation or correlation for some of the metrics describe.

References.

[1] T. Shafee, G. Masukume, L. Kipersztok, D. Das, M. Häggström, and J. Heilman, "Evolution of Wikipedia's medical content: past, present and future," *J Epidemiol Community Health*, p. jech-2016-208601, Aug. 2017.

[2] R. Al Tamime, R. Giordano, and W. Hall, "Humans and bots in controversial environments: A closer look at their interactions," *Qatar Found. Annu. Res. Conf. Proc.*, vol. 2018, no. 3, p. ICTPD425, Mar. 2018.

[3] A. Azzam *et al.*, "Why Medical Schools Should Embrace Wikipedia: Final-Year Medical Student Contributions to Wikipedia Articles for Academic Credit at One School," *Acad. Med.*, vol. 92, no. 2, pp. 194–200, Feb. 2017.

[4] J. M. Heilman and A. G. West, "Wikipedia and Medicine: Quantifying Readership, Editors, and the Significance of Natural Language," *J. Med. Internet Res.*, vol. 17, no. 3, Mar. 2015.

[5] R. Xu, L. Li, and Q. Wang, "Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature," *Bioinformatics*, vol. 29, no. 17, pp. 2186–2194, Sep. 2013.

[6] A. Rodríguez-González and G. Alor-Hernández, "An approach for solving multi-level diagnosis in high sensitivity medical diagnosis systems through the application of semantic technologies," *Comput. Biol. Med.*, vol. 43, no. 1, pp. 51–62, Jan. 2013.

[7] L.-W. Ku, W.-L. Li, and T.-C. Chang, "Disease Detection and Symptom Tracking by Retrieving Information from the Web.," 2013.

[8] K. S. Hickmann *et al.*, "Forecasting the 2013–2014 Influenza Season Using Wikipedia," *PLOS Comput. Biol.*, vol. 11, no. 5, p. e1004239, May 2015.

[9] N. Generous, G. Fairchild, A. Deshpande, S. Y. D. Valle, and R. Priedhorsky, "Global Disease Monitoring and Forecasting with Wikipedia," *PLOS Comput. Biol.*, vol. 10, no. 11, p. e1003892, Nov. 2014.

[10] V. Cozza, M. Petrocchi, and A. Spognardi, "A Matter of Words: NLP for Quality Evaluation of Wikipedia Medical Articles," in *Web Engineering*, 2016, pp. 448–456.

[11] Y. Ling, Y. An, and S. A. Hasan, "Improving Clinical Diagnosis Inference through Integration of Structured and Unstructured Knowledge," 2017.

[12] Eduardo P. García del Valle, Gerardo Lagunes García, Lucia Prieto Santamaria, Massimiliano Zanin, Alejandro Rodríguez González, and Ernestina Menasalvas Ruiz, "Evaluating Wikipedia as a source of information for disease understanding," presented at the 31st IEEE CBMS International Symposium on Computer-Based Medical Systems, 2018.

[13] V. Pedro, R. Stefan Niculescu, and L. Lita, "Okinet: Automatic Extraction of a Medical Ontology From Wikipedia," Jan. 2008.

[14] D. Milne, O. Medelyan, and I. H. Witten, "Mining Domain-Specific Thesauri from Wikipedia: A Case Study," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, USA, 2006, pp. 442–448.

[15] D. R. Schlegel, C. Crowner, and P. L. Elkin, "Automatically Expanding the Synonym Set of SNOMED CT using Wikipedia," *Stud. Health Technol. Inform.*, vol. 216, pp. 619–623, 2015.

[16] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," in *The Semantic Web*, Springer, Berlin, Heidelberg, 2007, pp. 722–735.

[17] A. Rodríguez-González, M. Martínez-Romero, R. Costumero, M. D. Wilkinson, and E. Menasalvas-Ruiz, "Diagnostic Knowledge Extraction from MedlinePlus: An Application for Infectious Diseases," in *9th International Conference on Practical Applications of Computational Biology and Bioinformatics*, Springer, Cham, 2015, pp. 79–87.

[18] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," *Proc. AMIA Symp.*, pp. 17–21, 2001.

[19] J. D. Osborne, B. Gyawali, and T. Solorio, "Evaluation of YTEX and MetaMap for clinical concept recognition," *ArXiv14021668 Cs*, Feb. 2014.

[20] P. Gooch and A. Roudsari, "A tool for enhancing MetaMap performance when annotating clinical guideline documents with UMLS concepts," presented at the IDAMAP Workshop at 13th Conference on Artificial Intelligence in Medicine (AIME'11), Bled, Slovenia, 2011.

[21] L. Al-Safadi, R. Alomran, and F. Almutairi, "Evaluation of Metamap Performance in Radiographic Images Retrieval," *Res. J. Appl. Sci. Eng. Technol.*, vol. 6, pp. 4231–4236, Dec. 2013.

[22] S. Meystre and P. J. Haug, "Evaluation of Medical Problem Extraction from Electronic Clinical Documents Using MetaMap Transfer (MMTx)," *Stud. Health Technol. Inform.*, vol. 116, pp. 823–828, 2005.

[23] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, "Human symptoms-disease network," *Nat. Commun.*, vol. 5, p. 4212, Jun. 2014.

[24] J. Amberger, C. Bocchini, and A. Hamosh, "A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®)," *Hum. Mutat.*, vol. 32, no. 5, pp. 564–567, May 2011.