

## Dataset description

---

# Wikipedia Knowledge Graph

*Wenceslao Arroyo-Machado*<sup>1</sup>, *Daniel Torres-Salinas*<sup>1</sup>, *Rodrigo Costas*<sup>2,3</sup>

<sup>1</sup> Department of Information and Communication Sciences, University of Granada, Granada, Spain

<sup>2</sup> Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands

<sup>3</sup> DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, Stellenbosch, South Africa

**Summary:** This document includes the description of the Wikipedia Knowledge Graph dataset. It is composed of 9 files in tsv format with data about the pages of the English edition of Wikipedia and main elements related to it.

## Content

<b>1. Introduction</b>	<b>1</b>
<b>2. Dataset summary</b>	<b>1</b>
<b>3. Dataset schema</b>	<b>2</b>
<b>4. File summary</b>	<b>3</b>
4.1. page	3
4.2. page_property	4
4.3. page_link	4
4.4. category	4
4.5. page_category	5
4.6. url	5
4.7. page_url	5
4.8. pub	6
4.9. page_pub	7

# 1. Introduction

Wikipedia is the largest and most read online free encyclopedia currently existing. As such, Wikipedia offers a large amount of data on all its own contents and interactions around them, as well as different types of open data sources. This makes Wikipedia a unique data source that can be analyzed with quantitative data science techniques. However, the enormous amount of data makes it difficult to have an overview, and sometimes many of the analytical possibilities that Wikipedia offers remain unknown. In order to reduce the complexity of identifying and collecting data on Wikipedia and expanding its analytical potential, after collecting different data from various sources and processing them, we have generated a dedicated Wikipedia Knowledge Graph aimed at facilitating the analysis, contextualization of the activity and relations of Wikipedia pages, in this case limited to its English edition. We share this Knowledge Graph dataset in an open way, aiming to be useful for a wide range of researchers, such as informetricians, sociologists or data scientists.

## 2. Dataset summary

There are a total of 9 files, all of them in tsv format, and they have been built under a relational structure. The main one that acts as the core of the dataset is the *page* file, after it there are 4 files with different entities related to the Wikipedia pages (*category*, *url*, *pub* and *page\_property* files) and 4 other files that act as "intermediate tables" making it possible to connect the pages both with the latter and between pages (*page\_category*, *page\_url*, *page\_pub* and *page\_link* files).

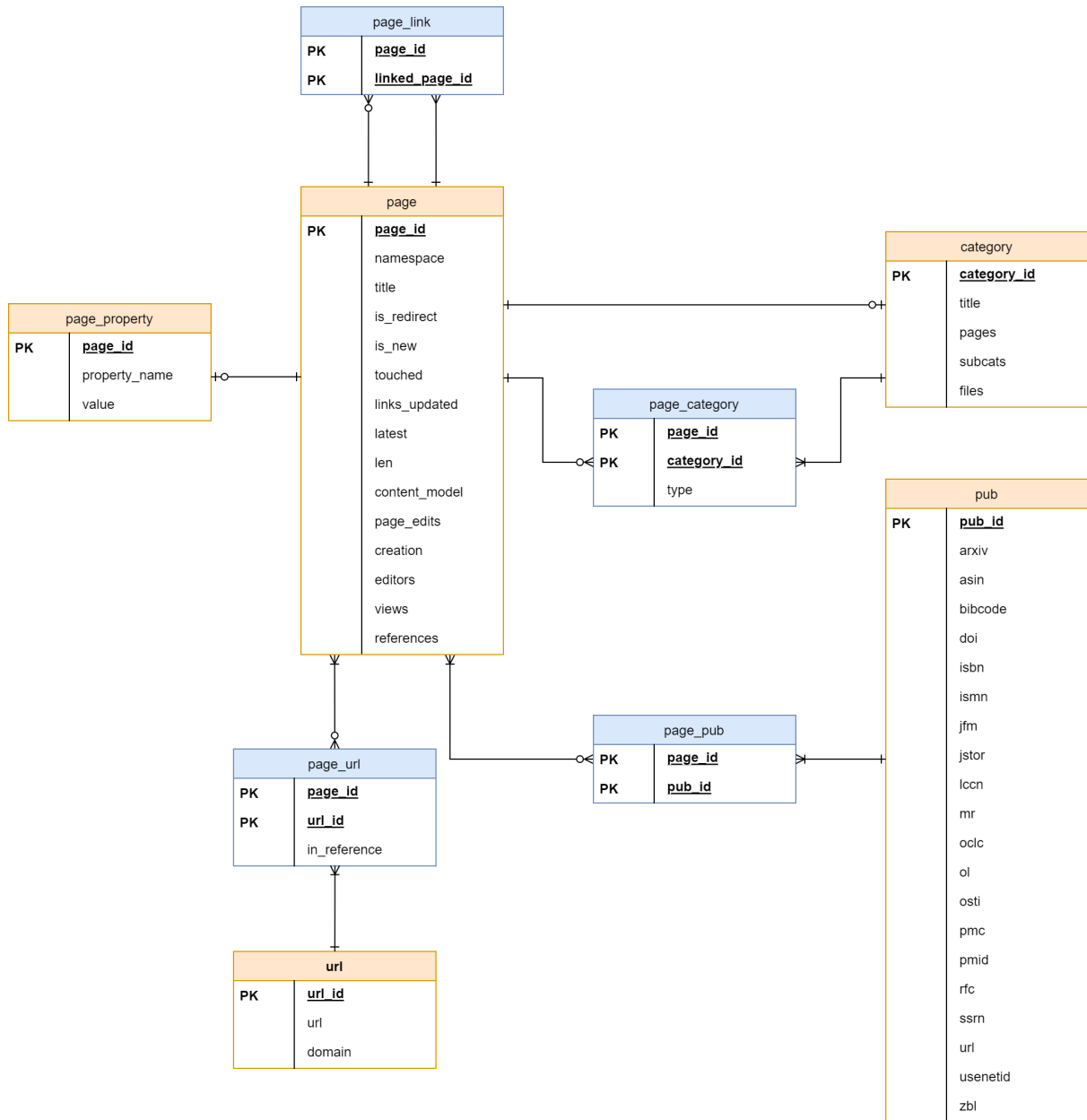
**Table 1.** Summary of the 9 files included in the dataset of Wikipedia Knowledge Graph.

Table	Up to date	Type	Dimensión	Size
<i>page</i>	07-01-2021	core	53,710,529 x 15	5.47 GB
<i>page_property</i>	07-01-2021	base	28,967,070 x 3	1.07 GB
<i>page_link</i>	-	intermediate	566,536,991 x 2	9.36 GB
<i>category</i>	07-01-2021	base	2,179,622 x 5	100 MB
<i>page_category</i>	-	intermediate	165,501,704 x 3	3.52 GB
<i>url</i>	05-2020 07-01-2021	base	51,923,982 x 3	4.03 GB
<i>page_url</i>	-	intermediate	65,554,992 x 3	1.23 GB
<i>pub</i>	05-2020	base	2,367,548 x 21	153 MB
<i>page_pub</i>	-	intermediate	3,728,522 x 2	59 MB

### 3. Dataset schema

Figure 1 shows the structure and relationships of the different data files following an Entity-Relationship Model (ER Modeling).


**Figure 1.** Entity-Relationship Diagram (ERD) of the Wikipedia Knowledge Graph dataset.



## 4. File summary

### 4.1. page



The main element of Wikipedia is the page, which can be an encyclopedic article, its discussion or a user, among other types. This file contains the main information about the Wikipedia pages.

Field	Type	Null
<b>page_id</b> 	<i>int</i>	<i>Not</i>
namespace	<i>int</i>	<i>Not</i>
title	<i>nvarchar(255)</i>	<i>Not</i>
is_redirect	<i>bit</i>	<i>Not</i>
is_new	<i>bit</i>	<i>Not</i>
touched	<i>char(14)</i>	<i>Not</i>
links_updated	<i>varchar(14)</i>	
latest	<i>int</i>	<i>Not</i>
len	<i>int</i>	<i>Not</i>
content_model	<i>varchar(32)</i>	
page_edits	<i>int</i>	
creation	<i>varchar(25)</i>	
editors	<i>int</i>	
views*	<i>int</i>	
references**	<i>int</i>	

\*The visits only cover the period from April 1 to June 30, 2021  
\*\*As of May 2021



## 4.2. page\_property

Additional metadata on Wikipedia pages among which are properties of specific types of pages, for example if a category is hidden, are included in this file.

Field	Type	Null
<b>page_id</b> 	<i>int</i>	<i>Not</i>
<b>property_name</b> 	<i>varchar(28)</i>	<i>Not</i>
property	<i>nvarchar(max)</i>	


## 4.3. page\_link

This file includes the links between Wikipedia pages.

Field	Type	Null
<b>page_id</b> 	<i>int</i>	<i>Not</i>
<b>linked_page_id</b> 	<i>int</i>	<i>Not</i>



## 4.4. category

This file includes all existing Wikipedia categories and some metadata about them. Wikipedia categories are tags used for the categorization of Wikipedia pages.

Field	Type	Null
<b>category_id</b> 	<i>int</i>	<i>Not</i>
title	<i>nvarchar(255)</i>	<i>Not</i>
pages	<i>int</i>	<i>Not</i>
subcats	<i>int</i>	<i>Not</i>
files	<i>int</i>	<i>Not</i>


## 4.5. page\_category

This file identifies which categories are associated with each Wikipedia page.

Field	Type	Null
<b>page_id</b> 	<i>int</i>	<i>Not</i>
<b>category_id</b> 	<i>int</i>	<i>Not</i>
type	<i>varchar(6)</i>	<i>Not</i>



## 4.6. url

All external links (URLs of websites outside Wikipedia) included in Wikipedia pages.

Field	Type	Null
<b>url_id</b> 	<i>int</i>	<i>Not</i>
url	<i>nvarchar(max)</i>	<i>Not</i>
domain	<i>nvarchar(max)</i>	<i>Not</i>


## 4.7. page\_url

Links between Wikipedia pages and external links.

Field	Type	Null
<b>page_id</b> 	<i>int</i>	<i>Not</i>
<b>url_id</b> 	<i>int</i>	<i>Not</i>
in_reference	<i>bit</i>	<i>Not</i>

## 4.8. pub

All publications referenced in Wikipedia pages. Publications are considered to be any material with an identifier (ISBN, DOI...), not only those of a scientific nature. It should be noted that all materials with a URL are not identified as publications. The URL is only included in addition.

Field	Type	Null
<b>pub_id</b> 	<i>int</i>	<i>Not</i>
arxiv	<i>nvarchar(22)</i>	
asin	<i>nvarchar(10)</i>	
bibcode	<i>nvarchar(21)</i>	
doi	<i>nvarchar(156)</i>	
isbn	<i>nvarchar(13)</i>	
ismn	<i>nvarchar(19)</i>	
jfm	<i>nvarchar(10)</i>	
jstor	<i>nvarchar(70)</i>	
lccn	<i>nvarchar(16)</i>	
mr	<i>nvarchar(8)</i>	
oclc	<i>nvarchar(13)</i>	
ol	<i>nvarchar(11)</i>	
osti	<i>nvarchar(27)</i>	
pmc	<i>nvarchar(10)</i>	
pmid	<i>int</i>	
rfc	<i>nvarchar(13)</i>	
ssrn	<i>int</i>	
url	<i>nvarchar(max)</i>	
usenetid	<i>nvarchar(68)</i>	
zbl	<i>nvarchar(10)</i>	

## 4.9. page\_pub

This file collects the links between Wikipedia pages and the publications they reference.

Field	Type	Null
<b>page_id</b> 🔑	<i>int</i>	<i>Not</i>
<b>pub_id</b> 🔑	<i>int</i>	<i>Not</i>