Detecting Wikipedia articles strongly based on single library collections

247 Dutch Wikipedia articles that wouldn't be here without Delpher and DBNL, with 33.000 views each month

Olaf Janssen, 21 May 2020

In this post I will illustrate an approach to detect Wikipedia articles whose contents are fully or largely based on content from a single online source, such as a full-text digitized newspaper archive or a digital text library. Using Dutch Wikipedia I'll track down 247 articles that owe their existence to Delpher and DBNL, two full-text collections operated by the KB, the national library of the Netherlands.

This approach might be relevant for GLAMs that have digital text collections used by the Wikipedia community for writing articles.

Three key players: Delpher, DBNL and KB

To understand the rest of this post, I'll start with a short introduction of three key players:

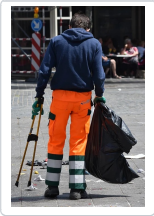
Delpher is a website containing over 100 million full-text digitized pages from Dutch historical newspapers, books and periodicals.

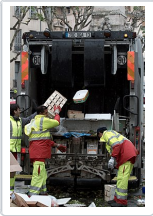
DBNL is the Digital Library for Dutch Literature (Dutch: *Digitale Bibliotheek voor de Nederlandse Letteren*, DBNL), a website about Dutch language and Dutch literature. It contains thousands of literary texts, secondary literature and additional information, like biographies, portrayals etcetera, and hyperlinks.

The Koninklijke Bibliotheek (KB) is the national library of the Netherlands. Both Delpher and DBNL are services operated by the KB.

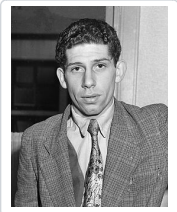
OK, let's go: Quiz time!

What is the connection between a garbage man, a garbage bag and a garbage truck?





Or between the Dutch soccer players Cor van der Gijp, Gerrie ter Horst and Joop van Daele?




Or between Hotel Des Indes and the International Press Museum, both located in The Hague, The Netherlands?

Or between a children's song book and the literary magazine 'Forum' (1932-1935)?


The answer:

The Dutch Wikipedia articles about these things probably wouldn't be there without Delpher or DBNL. In other words: the contents of these articles is fully or largly based on the contents of Delpher and/or DBNL. These articles owe their existence to the KB as the content supplier and the Wikipedia community piecing together all those pieces of Delpher/DBNL content into Wikipedia articles for millions of potential readers.

A more detailed look

Every two years I measure a number of indicators about the reach and reuse of KB collections via the Wikimedia platforms, most recently in February 2020. I would like to share one of the insights I gained from that analysis: *Dutch Wikipedia contains dozens of articles that would not have existed today without Delpher and/or DBNL.*

To be more specfific, last February I determined

- *Which articles* on Dutch Wikipedia contain one or more references (links, URLs) to websites of the KB, specifically to Delpher and DBNL. In other words: which articles are partially, largely or fully based on the content of KB websites (more details in Dutch)
- *How often* these articles are requested every month (more details in Dutch)
- *How many references* to KB websites all those articles contain (more details in Dutch). After all, one single article can contain multiple references. This is cleary illustrated in the article about Hotel Des Indes, which contains no fewer than 74 links to newspaper articles in Delpher.

Verwijzingen

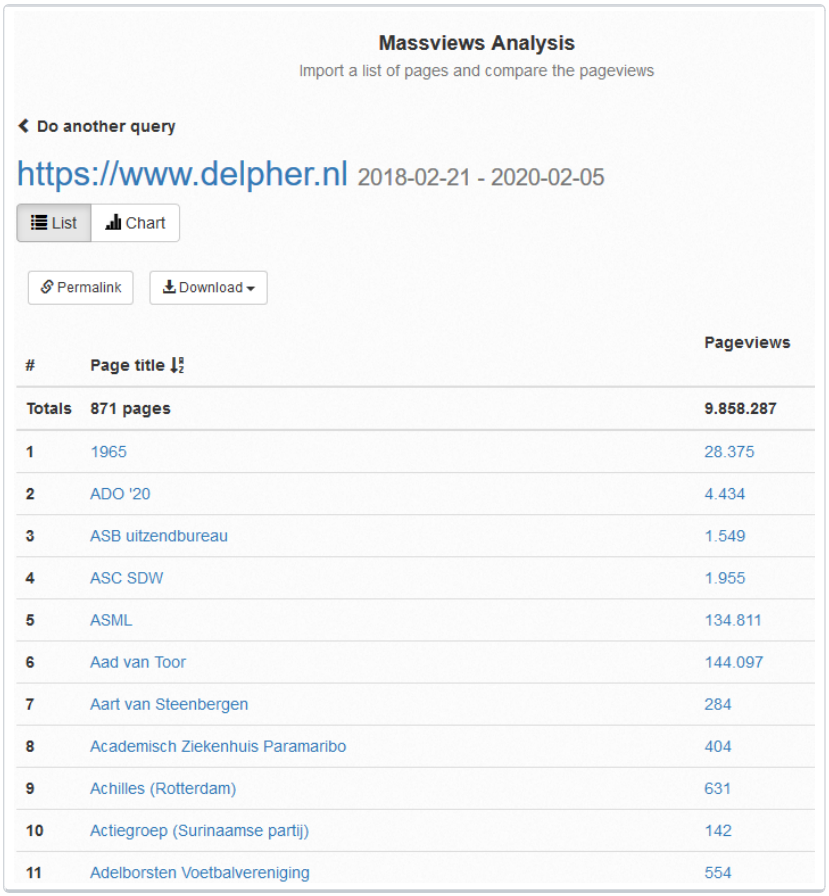
7. ↑ Koninklijke Bibliotheek/Delpher: De Tijd, 27-04-1881, pagina 2
8. ↑ Koninklijke Bibliotheek/Delpher: Haagsche Courant, 18-07-1887, pagina 2
9. ↑ Koninklijke Bibliotheek/Delpher: Het Nieuws van den Dag, 21-09-1882, pagina 7
10. ↑ Koninklijke Bibliotheek/Delpher: Provinciale Noordbrabantsche en 's Hertogenbossche Courant, 09-01-1883, pagina 1
11. ↑ Koninklijke Bibliotheek/Delpher: Algemeen Handelsblad, 07-06-1883, pagina 3
12. ↑ Koninklijke Bibliotheek/Delpher: Haagsche Courant, 18-07-1883, pagina 1
13. ↑ Koninklijke Bibliotheek/Delpher: Aankomst der Transvalers in Nederland, Algemeen Handelsblad, 29-02-1884, pagina 3
14. ↑ Koninklijke Bibliotheek/Delpher: Haagsche Courant, 16-06-1885, pagina 1
15. ↑ Koninklijke Bibliotheek/Delpher: Haagsche Courant, 16-07-1885, pagina 1
16. ↑ Koninklijke Bibliotheek/Delpher: Delftsche courant, 24-07-1888, pagina 1
17. ↑ Koninklijke Bibliotheek/Delpher: Algemeen Handelsblad, pagina 1, 14-07-1889
18. ↑ Haagsche Courant: Binnenland. (vervolg), 19-07-1890, pagina 3
19. ↑ Koninklijke Bibliotheek/Delpher: Haagsch Courant, 09-08-1890, pagina 1
20. ↑ Koninklijke Bibliotheek/Delpher: Het Nieuws van den Dag, 10-08-1894, pagina 11
21. ↑ Koninklijke Bibliotheek/Delpher: De Tijd, 29-11-1890, pagina 2
22. ↑ Koninklijke Bibliotheek/Delpher: Haagsche Courant, 17-07-1891, pagina 1
23. ↑ Koninklijke Bibliotheek/Delpher: De Tijd, 29-07-1892, pagina 2
24. ↑ Koninklijke Bibliotheek/Delpher: De Telegraaf, 02-09-1898, pagina 1
25. ↑ Koninklijke Bibliotheek/Delpher: De Tijd, 23-08-1892, pagina 2
26. ↑ Koninklijke Bibliotheek/Delpher: Algemeen Handelsblad, 14-09-1895, pagina 3
27. ↑ Koninklijke Bibliotheek/Delpher: De Tijd, 08-07-1896, pagina 2
36. ↑ Koninklijke Bibliotheek/Delpher: De Tijd, 19-11-1906, pagina 2
37. ↑ Koninklijke Bibliotheek/Delpher: Delftsche Courant, 08-04-1909, pagina 1
38. ↑ Koninklijke Bibliotheek/Delpher: Haagsche Courant, 03-06-1909, pagina 5
39. ↑ Koninklijke Bibliotheek/Delpher: Roosevelt in Den Haag, Algemeen Dagblad, 30-04-1910, pagina 2
40. ↑ Koninklijke Bibliotheek/Delpher: Het Nieuws van den Dag, 21-04-1911, pagina 5
41. ↑ Koninklijke Bibliotheek/Delpher: Haagsche Courant, 07-07-1911, pagina 5
42. ↑ Koninklijke Bibliotheek/Delpher: R.A. Taft, Provinciale Geldersche en Nijmeegsche Courant, 22-07-1911, pagina 1
43. ↑ Koninklijke Bibliotheek/Delpher: Haagsche Courant, 12-08-1912, pagina 1
44. ↑ Koninklijke Bibliotheek/Delpher: Het Nieuws van den Dag, 12-09-1913, pagina 12
45. ↑ Koninklijke Bibliotheek/Delpher: Algemeen Handelsblad, 14-06-1921, pagina 2
46. ↑ Haagsche Courant: Japansch bezoek aan onze stad., 28-04-1925, pagina 13
47. ↑ Koninklijke Bibliotheek/Delpher: De Telegraaf, 12-10-1926, pagina 5
48. ↑ Koninklijke Bibliotheek/Delpher: Het Vaderland, 17-05-1932, pagina 1
49. ↑ Koninklijke Bibliotheek/Delpher: De president van Liberia, Provinciale Noordbrabantsche en 's Hertogenbossche Courant, 28-09-1927, pagina 8
50. ↑ Koninklijke Bibliotheek/Delpher: Japansch bezoek, Leeuwarder Courant, 26-10-1927, pagina 1
51. ↑ Koninklijke Bibliotheek/Delpher: De Graafschap-bode, 26-10-1928, pagina 2
52. ↑ Koninklijke Bibliotheek/Delpher: De Telegraaf, 26-10-1929, pagina 12
53. ↑ Koninklijke Bibliotheek/Delpher: Haagsche Courant, 31-10-1930, pagina 5
54. ↑ Koninklijke Bibliotheek/Delpher: Algemeen Handelsblad, 14-02-1938, pagina 12

Approach in 4 steps

During this measurement process I started to notice that there are quite a few *Hotel Des Indes*-like articles: articles containing a striking amount of links to Delpher and/or DBNL. That triggered my curiosity, so I went deeper and more systematic, in 4 steps.

Step 1: article lists

I started out by making an overview of all articles on Dutch Wikipedia containing one or more links to Delpher or DBNL. I did this using the Massviews Analysis tool, which takes a URL (or rather: a URL pattern, or base-URL) as input, and returns a list of articles containing that URL pattern. The screenshot below is based on the URL *https://www.delpher.nl* (click for live tool, might take some time)



Massviews Analysis

Import a list of pages and compare the pageviews

Do another query

https://www.delpher.nl 2018-02-21 - 2020-02-05

List | Chart

Permalink | Download

#	Page title	Pageviews
Totals	871 pages	9.858.287
1	1965	28.375
2	ADO '20	4.434
3	ASB uitzendbureau	1.549
4	ASC SDW	1.955
5	ASML	134.811
6	Aad van Toor	144.097
7	Aart van Steenbergen	284
8	Academisch Ziekenhuis Paramaribo	404
9	Achilles (Rotterdam)	631
10	Actiegroep (Surinaamse partij)	142
11	Adelborsten Voetbalvereniging	554

I used this tool for all Delpher URLs (don't forget the persistent KB-resolver base-URLs such as http://resolver.kb.nl/resolve?urn=ddd, see column 3 of this table for all base-URLs). I merged and de-duplicated the resulting article lists, and converted the outcome to Excel, the final result is a list of approx. 6.800 articles containing one or more Delpher URLs.

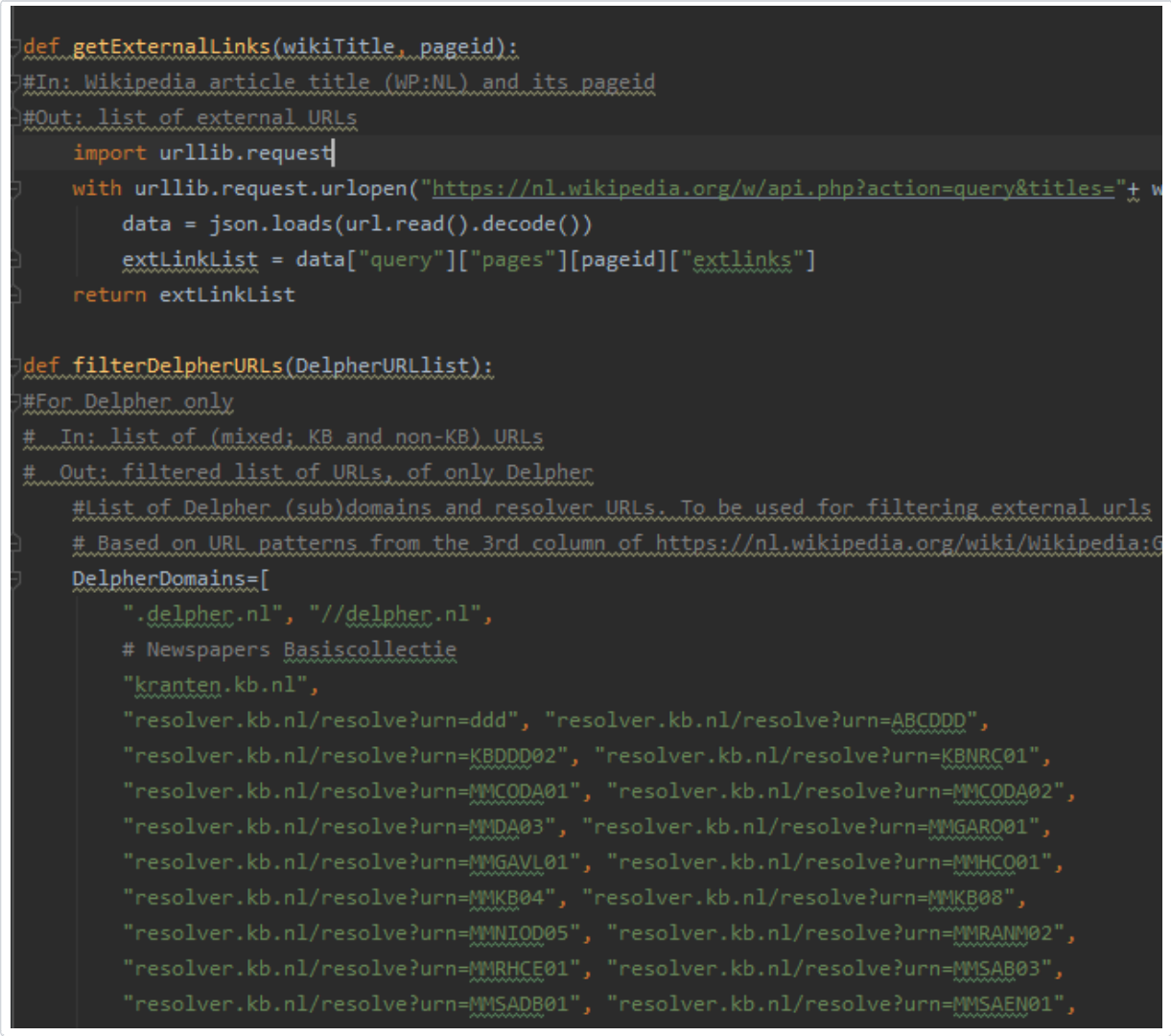
2	https://nl.wikipedia.org/wiki/...die_Revolutie_niet_begrepen!...
3	https://nl.wikipedia.org/wiki/10_jaar_Bassie_&_Adriaan
4	https://nl.wikipedia.org/wiki/10_juli
5	https://nl.wikipedia.org/wiki/12-verdiepingenhuis
6	https://nl.wikipedia.org/wiki/13_november
7	https://nl.wikipedia.org/wiki/17_maart
8	https://nl.wikipedia.org/wiki/19_juni
9	https://nl.wikipedia.org/wiki/1965
10	https://nl.wikipedia.org/wiki/2_mei
11	https://nl.wikipedia.org/wiki/24_december
12	https://nl.wikipedia.org/wiki/2Amsterdam

I used a similar workflow for DBNL (URL patterns *http(s)://*.dbnl.org*), resulting in a list of just over 7.600 unique Wikipedia articles.

Step 2: external links

Once I had those article lists, for each article I determined which (and how many) external links it contains, and which of those links point to Delpher (or DBNL). I did this using the MediaWiki API and

Python script (for Delpher and for DBNL). In the screenshot below of the Delpher script you can see that filtering is done on the resolver base-URLs of the Delpher Newspapers subset.



```
def getExternalLinks(wikiTitle, pageid):
#In: Wikipedia article title (WP:NL) and its pageid
#Out: list of external URLs
    import urllib.request
    with urllib.request.urlopen("https://nl.wikipedia.org/w/api.php?action=query&titles="+ w
        data = json.loads(url.read().decode())
        extLinkList = data["query"]["pages"][pageid]["extlinks"]
    return extLinkList

def filterDelpherURLs(DelpherURLlist):
#For Delpher only
#  In: list of (mixed; KB and non-KB) URLs
#  Out: filtered list of URLs, of only Delpher
    #List of Delpher (sub)domains and resolver URLs. To be used for filtering external urls
    # Based on URL patterns from the 3rd column of https://nl.wikipedia.org/wiki/Wikipedia:G
    DelpherDomains=[
        ".delpher.nl", "//delpher.nl",
        # Newspapers Basiscollectie
        "kranten.kb.nl",
        "resolver.kb.nl/resolve?urn=ddd", "resolver.kb.nl/resolve?urn=ABCDDD",
        "resolver.kb.nl/resolve?urn=KBDDD02", "resolver.kb.nl/resolve?urn=KBNRC01",
        "resolver.kb.nl/resolve?urn=MMCODA01", "resolver.kb.nl/resolve?urn=MMCODA02",
        "resolver.kb.nl/resolve?urn=MMDA03", "resolver.kb.nl/resolve?urn=MMGARO01",
        "resolver.kb.nl/resolve?urn=MMGAVL01", "resolver.kb.nl/resolve?urn=MMHCO01",
        "resolver.kb.nl/resolve?urn=MMKB04", "resolver.kb.nl/resolve?urn=MMKB08",
        "resolver.kb.nl/resolve?urn=MMNIOD05", "resolver.kb.nl/resolve?urn=MMRANM02",
        "resolver.kb.nl/resolve?urn=MMRHCE01", "resolver.kb.nl/resolve?urn=MMSAB03",
        "resolver.kb.nl/resolve?urn=MMSADB01", "resolver.kb.nl/resolve?urn=MMSAEN01",
```

This step eventually yields an Excel that (for Delpher) looks like this:

WikiURL		NrOfExtLinks	NrOfDelpherLinks
https://nl.wikipedia.org/wiki/...die_Revolutie_niet_begrepen!...	Klik	16	9
https://nl.wikipedia.org/wiki/10_jaar_Bassie_&_Adriaan	Klik	5	3
https://nl.wikipedia.org/wiki/10_juli	Klik	2	1
https://nl.wikipedia.org/wiki/12-verdiepingenhuis	Klik	10	3
https://nl.wikipedia.org/wiki/13_november	Klik	2	1
https://nl.wikipedia.org/wiki/17_maart	Klik	2	1
https://nl.wikipedia.org/wiki/19_juni	Klik	1	1
https://nl.wikipedia.org/wiki/1965	Klik	3	1
https://nl.wikipedia.org/wiki/2_mei	Klik	2	1
https://nl.wikipedia.org/wiki/24_december	Klik	2	1
https://nl.wikipedia.org/wiki/2Amsterdam	Klik	10	2
https://nl.wikipedia.org/wiki/50_Kamers	Klik	4	1

For example, the first article "...die_Revolutie_niet_begrepen!..." contains 16 external links, 9 of which point to Delpher.

Step 3: link ratio

Because we are looking for articles that are entirely or largely based on contents from Delpher (or DBNL), it is useful to look at the so-called *link ratio* . That is the ratio of the total number of external links, and the number of those that link to Delpher. A link ratio of 1.00 means that *all* external links in an article are Delpher links. The lower the link ratio, the smaller the relative number of Delpher links in the article.

WikiURL		NrOfExtLinks	NrOfDelpherLinks	LinkRatio
https://nl.wikipedia.org/wiki/...die_Revolutie_niet_begrepen!...	Klik	16	9	0,563
https://nl.wikipedia.org/wiki/10_jaar_Bassie_&_Adriaan	Klik	5	3	0,600
https://nl.wikipedia.org/wiki/10_juli	Klik	2	1	0,500
https://nl.wikipedia.org/wiki/12-verdiepingenhuis	Klik	10	3	0,300
https://nl.wikipedia.org/wiki/13_november	Klik	2	1	0,500
https://nl.wikipedia.org/wiki/17_maart	Klik	2	1	0,500
https://nl.wikipedia.org/wiki/19_juni	Klik	1	1	1,000
https://nl.wikipedia.org/wiki/1965	Klik	3	1	0,333
https://nl.wikipedia.org/wiki/2_mei	Klik	2	1	0,500
https://nl.wikipedia.org/wiki/24_december	Klik	2	1	0,500
https://nl.wikipedia.org/wiki/2Amsterdam	Klik	10	2	0,200
https://nl.wikipedia.org/wiki/50_Kamers	Klik	4	1	0,250
https://nl.wikipedia.org/wiki/8_juli	Klik	3	1	0,333

Step 4: threshold criteria

Next, to determine whether an article owes its existence largely to Delpher (or DBNL), I use two threshold criteria:

1. The article must contain a minimum number of external links, as its content must be sufficiently based on external sources.
2. The link ratio must exceed a certain threshold in order to have Delpher (or DBNL) as an external source sufficiently often.

There is some freedom in the choice of both thresholds, I have used the following:

- for Delpher: Number of external links >=6, link ratio>=0.75
- for DBNL: Number of external links >=4, link ratio>=0.7

This results in the following table for Delpher

WikiURL		NrOfExtLinks	NrOfDel	LinkRatio
https://nl.wikipedia.org/wiki/Vuilnisman	Klik	30	30	1,000
https://nl.wikipedia.org/wiki/Lijst_van_burgemeesters_van_Aengwirden	Klik	21	21	1,000
https://nl.wikipedia.org/wiki/Lijst_van_burgemeesters_van_Zaamslag	Klik	19	19	1,000
https://nl.wikipedia.org/wiki/AFC_Ajax_in_het_seizoen_1911/12	Klik	17	17	1,000
https://nl.wikipedia.org/wiki/Cor_van_der_Gijp	Klik	16	16	1,000
https://nl.wikipedia.org/wiki/Executie_van_Adriaan_de_Klerk_en_Cornelis_de_Jong	Klik	12	12	1,000
https://nl.wikipedia.org/wiki/Hendrik_Croes	Klik	11	11	1,000
https://nl.wikipedia.org/wiki/Onze-Lieve-Vrouw-Geboortekerk_(Hoogmade,_1875)	Klik	11	11	1,000
https://nl.wikipedia.org/wiki/Freek_van_der_Gijp	Klik	10	10	1,000
https://nl.wikipedia.org/wiki/Salon_van_de_Maassteden	Klik	10	10	1,000
https://nl.wikipedia.org/wiki/Theo_van_Hengel	Klik	31	27	0,871
https://nl.wikipedia.org/wiki/Hotel_Des_Indes_(Den_Haag)	Klik	85	74	0,871
https://nl.wikipedia.org/wiki/Sijtje_Boes	Klik	14	12	0,857
https://nl.wikipedia.org/wiki/Wim_van_Lent	Klik	14	12	0,857
https://nl.wikipedia.org/wiki/A.E._Thierens	Klik	7	6	0,857
https://nl.wikipedia.org/wiki/Eredivisie_(handbal)_1991-92	Klik	7	6	0,857
https://nl.wikipedia.org/wiki/Hendrik_Wielinga	Klik	7	6	0,857
https://nl.wikipedia.org/wiki/Henri_Antoine_Termijtelen	Klik	7	6	0,857

Analysis

The articles found in this way are **places where strong aggregation and republication of Delpher content takes place.** In other words: *These articles bring together information from Delpher related to people, places, events and other topics for a wide audience, as 80% of the Netherlands reads Wikipedia.* The same goes for DBNL.

If you look at the lists of the 'aggregation articles' obtained in this way, you see

For Delpher

- 193 articles owe their existence largely or fully to Delpher.
- The article Lijst van historische Nederlandse netnummers holds most Delpher links, 165 out of the 195 external links, with the above *Hotel Des Indes* coming second.

WikiURL		NrOfExtLinks	NrOfDelpherLinks	LinkRatio
https://nl.wikipedia.org/wiki/Lijst_van_historische_Nederlandse_netnummers	Klik	195	165	0,846
https://nl.wikipedia.org/wiki/Hotel_Des_Indes_(Den_Haag)	Klik	85	74	0,871
https://nl.wikipedia.org/wiki/Toon_van_den_Enden	Klik	51	50	0,980
https://nl.wikipedia.org/wiki/Zaro_Agha	Klik	44	43	0,977
https://nl.wikipedia.org/wiki/Lijst_van_burgemeesters_van_Landsmeer	Klik	38	31	0,816
https://nl.wikipedia.org/wiki/Lijst_van_burgemeesters_van_Twisk	Klik	33	25	0,758
https://nl.wikipedia.org/wiki/Jos._E._Vogt	Klik	32	30	0,938
https://nl.wikipedia.org/wiki/Groninger_Museum	Klik	32	24	0,750
https://nl.wikipedia.org/wiki/Theo_van_Hengel	Klik	31	27	0,871
https://nl.wikipedia.org/wiki/Vuilnisman	Klik	30	30	1,000

- The subject width of articles using Delpher as their main source is very large: from the garbage industry to luxury hotels, from politicians to people condemned to death and from music awards to Michelin-starred restaurants.
- Quite a few articles about sports - e.g. soccer players, annual overviews of swimming championships and korfball - heavily rely on Delpher, similar to articles listing mayors (*burgemeesters*) of Dutch towns and villages.

https://nl.wikipedia.org/wiki/Lijst_van_burgemeesters_van_Abbekerk	Klik	19	15	0,789
https://nl.wikipedia.org/wiki/Lijst_van_burgemeesters_van_Aengwirden	Klik	21	21	1,000
https://nl.wikipedia.org/wiki/Lijst_van_burgemeesters_van_Ameide	Klik	8	8	1,000
https://nl.wikipedia.org/wiki/Lijst_van_burgemeesters_van_Appingedam	Klik	7	6	0,857
https://nl.wikipedia.org/wiki/Lijst_van_burgemeesters_van_Est_en_Opijnen	Klik	20	15	0,750
https://nl.wikipedia.org/wiki/Lijst_van_burgemeesters_van_Gameren	Klik	8	6	0,750
https://nl.wikipedia.org/wiki/Lijst_van_burgemeesters_van_Hulsberg	Klik	13	11	0,846
https://nl.wikipedia.org/wiki/Lijst_van_burgemeesters_van_Landsmeer	Klik	38	31	0,816

For DBNL

- 54 articles owe their existence largely or fully to DBNL.
- Joost van den Vondel contains the most DBNL links, 32 out of 44 in total.

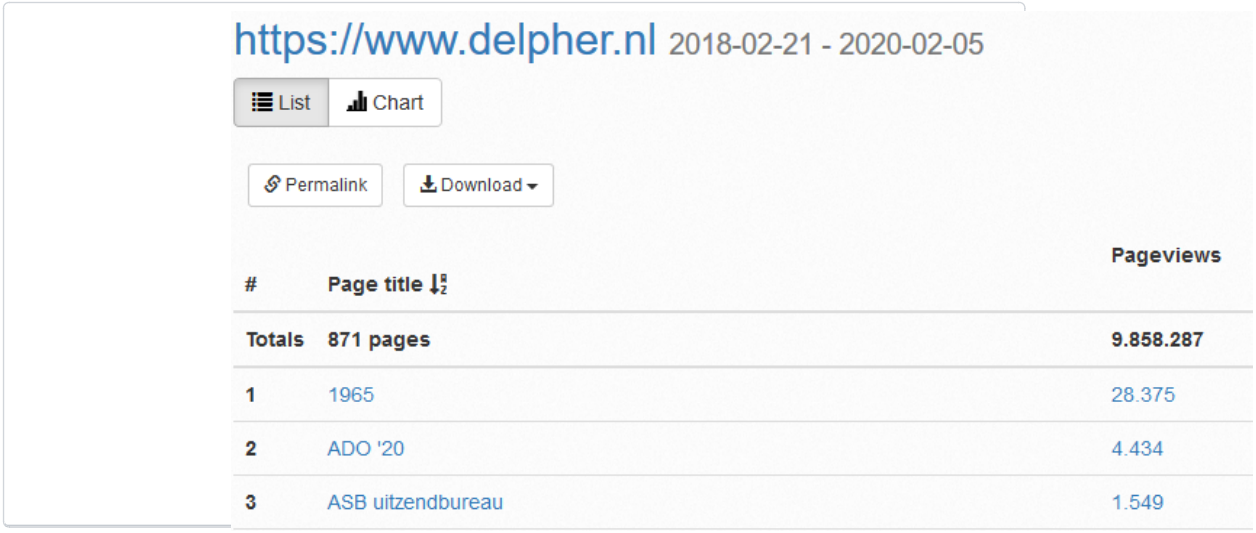
WikiURL		NrOfExtLinks	NrOfDBNLinks	LinkRatio
https://nl.wikipedia.org/wiki/Joost_van_den_Vondel	Klik	44	32	0,727
https://nl.wikipedia.org/wiki/Liedboek	Klik	26	24	0,923
https://nl.wikipedia.org/wiki/Kinderliedboek	Klik	18	15	0,833
https://nl.wikipedia.org/wiki/Jozef_van_Walleghem	Klik	14	14	1,000
https://nl.wikipedia.org/wiki/Surinaamse_literatuur	Klik	17	12	0,706
https://nl.wikipedia.org/wiki/Adam_in_ballingschap	Klik	15	12	0,800
https://nl.wikipedia.org/wiki/Faëton	Klik	9	9	1,000
https://nl.wikipedia.org/wiki/Piet_Schipperus	Klik	11	8	0,727

- In particular articles related to Dutch literature, writers, poets, publishers, books etc. use DBNL as their main source. The subject width of DBNL-based articles is much smaller than those based on Delpher. But this is hardly a surprise, given the contents and theme of DBNL vs. Delpher.

33.000 views every month

All very well these Wikpedia articles heavily based on Delpher and/or DBNL, but are they actually read by the public? I also looked into that.

For each article, the Massviews Analysis tool mentioned above also gives the number of requests (see the Pageviews column) during a certain period, in this case it is (almost) 2 years, from 21 Febr 2018 to 5 Febr 2020.



This allows us to determine the total number of requests for these 193 Delpher and 54 DBNL

aggregation articles during those two years.

- For Delpher: 343.821 page views
- For DBNL: 445.713 page views

In total, this amounts to 789.534 page views in 2 years, or an average of **33.000 requests per month.**

Raw data

The approach described above is also explained on Dutch Wikipedia. The Excels from which the above screenshots were created are available here on Github:

- List of Delpher aggregation articles
- List of DBNL aggregation articles

About the author

Olaf Janssen is the Wikimedia coordinator of the KB, the national library of the Netherlands. He contributes to Wikipedia, Wikimedia Commons and Wikidata as User:OlafJanssen

Reusing this article