Cybergeo : European Journal of Geography

Data papers
2020
928

One century of information diffusion in the Netherlands derived from a massive digital archive of historical newspapers: the DIGGER dataset

Un siècle de diffusion de l'information aux Pays-Bas extrait d'une archive de journaux historiques numérisés : la base de données DIGGER

Antoine Peris, Willem Jan Faber, Evert Meijers et Maarten van Ham

Résumés

English Français
Previous studies have highlighted the importance of having long term data for the study of cities, but such sources are relatively scarce. This is especially the case for data about relations between cities, which is a crucial aspect of urban dynamics. Over the last two decades, many efforts have been made to digitalize texts, including books and newspapers, which are primary sources on most of our societies. Researchers have shown that these massive digital archives can be used to identify macroscopic trends related to historical and cultural changes. The wealth of geographic information in such digital archives has not been used much, while they are very valuable for the study of cities. In this paper, we present DIGGER, a newly developed dataset that we built on Delpher, the digital archive of historical newspapers of the National Library of the Netherlands, by extracting geographical information from a selection of 102 million of news items. This dataset allowed us to study the spatial diffusion of information on and between the Dutch cities from a corpus of 81 newspapers published in 29 different cities between 1869 and 1994. This paper

presents the method developed to build the dataset as well as the validation steps for the accuracy of the place name recognition. This dataset can be used to study the evolution of the Dutch urban system as well as aspects related to the spatial diffusion of information and geographical bias in media coverage.

Les données couvrant de longues périodes temporelles sont relativement rares pour l'étude des villes et pourtant essentielles à la compréhension du temps long de leurs dynamiques. Ce problème est prégnant pour les données sur les relations interurbaines, à l'échelle des systèmes de ville. Au cours des deux dernières décennies, d'importants efforts de numérisation de textes anciens ont été entrepris, notamment de livres et de journaux qui constituent des sources très riches sur les sociétés qui les ont produites. Des chercheurs ont récemment montré que ces archives numériques massives peuvent être utilisées pour identifier des tendances macroscopiques en rapport avec des changements historiques et culturels. En revanche, peu d'études se sont intéressées à la richesse de l'information géographique qui peut être extraite de ces archives. Dans cet article, nous présentons DIGGER, une base de données construite à partir de Delpher, l'archive de journaux historiques numérisés de la Bibliothèque Nationale des Pays-Bas. Cette base a été construite suite à l'analyse du contenu de 102 millions d'articles et petites annonces publiés dans 81 journaux locaux de 29 villes néerlandaises dont la publication s'étale de 1869 à 1994. Nous présentons ici les différentes étapes nécessaires à la constitution de la base de données ainsi que la validation de notre algorithme identifiant les noms de lieux. Cette base de données peut être utilisée pour analyser plus d'un siècle de développement du système urbain des Pays-Bas ainsi que pour l'étude de la diffusion des informations ou des biais spatiaux dans la couverture médiatique.

Entrées d'index

Mots-clés : système de villes, flux, diffusion, histoire, bases de données
Keywords : system of cities, flows, diffusion, history, database

Texte intégral

Background & Summary

1 We have designed DIGGER in order to study the evolution of the Dutch urban system by investigating information flows extracted from historical newspapers that go back to 1869. Newspapers are full of geographical information as most of news items include one or more place names. However, studying this geographical information systematically is not an easy task. In this project, we have geocoded place names contained in a selection of 102 million news items to build origin-destination matrices with places mentioned in the news items (o) and places where the newspapers were issued (d) for 125 years (t). It takes the form of a cube with 3 dimensions: origin, destination and time.

2 Information circulation has been identified as a key factor in urban dynamics. Classical urban literature has highlighted the importance of available information on locational decisions of individuals, groups and firms and of its role as prerequisite for other kinds of people and goods movements. An early paper of Zipf (1946) used local newspapers to study the interactions between distant 'communities' and used this data in a gravity model. Allan Pred (1977) also used local newspapers from different American cities to measure the time it took for information to travel from one place to another. But because of the time and workforce needed for the data collection, these studies were limited to a very small number of cities or short periods of time.

3 However, with the recent development of computing techniques, it is now possible to upscale and systematize data collection from newspapers to analyse the information

circulation at the level of an entire territory. Over the last 10 years, interdisciplinary teams of computer scientists and humanity scholars have worked on extracting patterns from massive textual corpora illustrating historical and cultural changes. A seminal study by Michel et al. (2011) showed the potential of this approach by compiling 5 million digitalized books to provide quantitative insights on the evolution of grammar, as well as the detection of events such as pandemics, the influence of certain thinkers, or the evolution of gender bias in vocabulary. While the importance of such an approach was widely acknowledged, the study received a number of critiques related to the book selection (Morse-Gagné, 2011), and the fact that it did not include newspapers, which were thought to better reflect their time due to the frequency of publication (Schwartz, 2011). Indeed, the written press was the primary source to access information from distant places in most of the industrial societies for a long period of time. More recently, a study on British newspapers has used more refined techniques such as Named-Entity recognition to study the content of a massive corpus of historical newspapers (Lansdall-Welfare et al., 2017). While this study could look more precisely at historical and cultural trends, the analysis of the geographical focus, which was not the core of the study, remained at the stage of visualisation. However, problems related to extracting spatial information from text where not addressed, including the variety of scales (an article can mention a street, a city, a country, etc.) and ambiguities in place names.

4 Over the past few years, researchers have been increasingly interested in extracting geographical information from unstructured or semi-structured text data (Grasland, 2019; Meijers, Peris, 2018; Tranos, Kefalas, 2018). Extraction of spatial information from text was also carried out to map and analyse the global scientific production with a bibliometric database (Maisonobe, Eckert, Grossetti, Jégou, Milard, 2016). Mining these huge amounts of textual data is an important challenge for social sciences because these textual sources contain much information on social and economic processes, which are very often tied to places.

5 The motivation for the collection of DIGGER is to build a dataset on city-to-city interactions in order to test hypotheses related to the evolution of an urban system through history through the prism of information flows. Our objective is to look for macroscopic spatial trends in the way information is diffused and how this is changing over time. In the study related to this data paper, we investigate the changing role of distance, city size, cultural and administrative borders on the circulation of information on a sample of 31 local newspapers between 1869 and 1930 (Peris, Meijers, van Ham, *submitted*). We find evidences of a space-time contraction, with faraway places being increasingly covered. The changes in patterns of information flows are also characterized by a hierarchical selection process. Almost all newspapers report more and more about the 4 main cities of the Randstad (Amsterdam, Rotterdam, The Hague and Utrecht), at the expense of the intermediate provincial cities located close-by.

6 The following section presents aspects related to the methods used for the data collection. It focuses first on the selection of a corpus of newspapers from the digital archive and of a set of cities for which data is collected. Then, it presents issues in place names recognition and choices to deal with these issues. Afterwards, we present the tests that we did to have statistics on the accuracy of our method. We then present an example of use of the dataset by mapping the information field of different cities with information flows in 1871. Finally, we give some concluding remarks on the potential of this database and its possible uses.

Methods

Corpus selection

7 The first important step in any quantitative study using a text archive is to select a relevant corpus. The content of a digital archive might be influenced by many factors such as digitalization policies, projects targeting a specific part of the media landscape (a newspaper, a region or a time period) or copyrights issues. Carefully selecting the corpus can significantly reduce bias, and is necessary to create a dataset as representative as possible depending on the research question. We have applied four criteria in the selection of newspapers:

- the newspaper had to be issued after 1869;
- its publication place had to be in the Netherlands;
- the newspaper had to exist during at least two consecutive decades;
- and we dropped the many small newspapers that were published only during the Second World War.

8 The following paragraphs detail and justify the choices that have been made to select the final corpus of 81 newspapers.

9 We started by analysing the temporal coverage of Delpher. At the time the data collection started, there were 1970 different titles in the archive. Table 1 shows that most of the newspapers have a lifespan of 5 years or less. The very short lifespan of most of titles is consistent with the findings of Van Kranenburg et al. (1998) that show that during the period 1848-1997, most Dutch daily newspapers did not even survive a decade. Indeed, like any other firm, if a newspaper does not find a market (here a readership), it is most likely to disappear. The fact that some newspapers were able to survive during long periods of time is a proof that they were supported by a sufficiently large readership. Using the lifespan of newspapers is a crude but relatively reliable proxy for their importance. By selecting newspapers according to this time dimension, we ensure that they had a sufficient diffusion to stay alive at least two decades consecutively. This huge variability in duration is also reflected by the amount of news items published by the different newspapers.

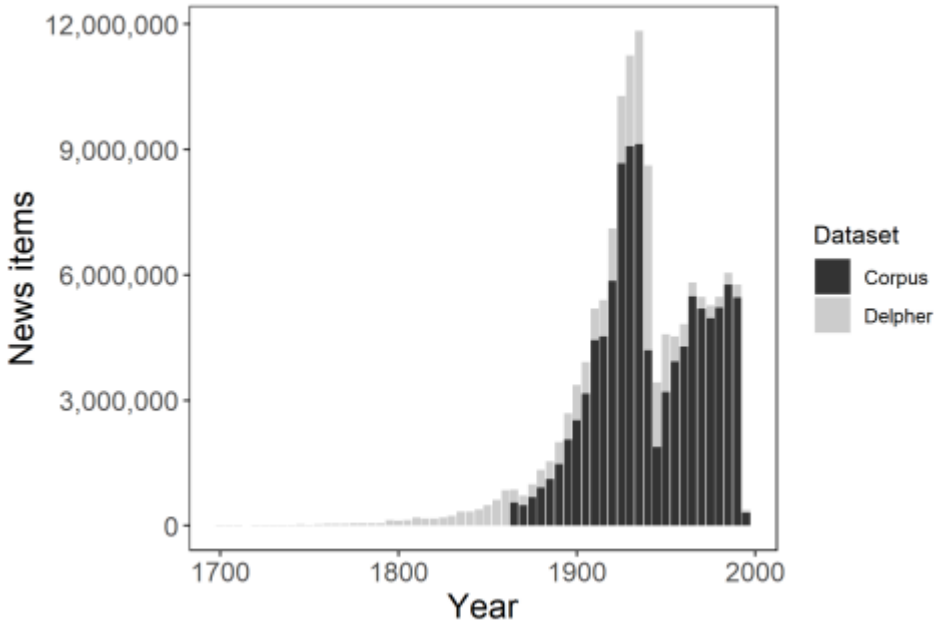
Table 1: Summary statistics of the Delpher corpus

	Min.	Median	Mean	Max.	St. dev.
Life span (Delpher)	5	5	11.14	180	17.69
Published items (Delpher)	1	123	67,843	13,307,273	491,679.6
Lifespan (selection)	11	50.50	53.39	129	28.53
Published items (selection)	1316	744,390	1,530,737	13,307,273	2,165,452

10 There are also important fluctuations in terms of number of publication of news items across the three centuries that are covered by the database (Figure 1). While the period 1700-1800 is characterized by a relatively small number, one can see an increase after the middle of the 19th century, followed by a very rapid rise and a peak that culminates in the 1940s. This tendency reflect the history of the Dutch press. The increase in the second half of the 19th can be explained by the abolishment of a tax on newspapers –

the '*dagbladzegel*' – that made them cheaper and affordable for a wider public. As we are interested by the amount of non-local information received by urban dwellers, we decided to take this time mark as our starting point, because from this period, newspapers became the backbone of information diffusion in the Netherlands. The following period is a period of development of the press, that ends in a peak during the Second World War, a period were many anti- and pro-German newspapers were created, most of the anti-German being underground. This resulted in the presence of a lot of short lived newspapers only published during the Second World War (n=2139) that can be very interesting for historians interested in the war but less relevant for long term studies. The application of these 4 criteria resulted in a sub-corpus of 81 newspapers that still cover an important part of the Delpher archive.

Figure 1: News items per year in Delpher and in the sub-corpus



Selection of cities

11 In any study on cities, the selection of entities to work on is never a trivial operation. The city is indeed a very complex object and its definition varies among countries and epochs. In our case, defining our primary units of analysis is made difficult by the fact that the data collection is meant for a corpus that covers more than one century.

12 Cities can be defined according to many criteria, they can be continuous build-up areas, functional entities, designated by a certain level of urban functions or by administrative status. A definition that is often used is municipal entities above a certain threshold of population. Because we are interested in identifying cities in texts, we must go beyond these definitions and identify the terms that relate to cities in the common language. We adopt what Goodchild and Li (2011) call a "placial" perspective. These authors make a distinction between the *spatial perspective*, where the geographic information is organized by coordinate systems, and the *placial perspective*, that focusses on places as "named domains in human discourse". For them, one is not more important than the other as "the name people give to places and points of interest constitute a very significant form of geographical information". We decided to look at the terms people use to say where they live because place names have a stronger inertia than the boundaries of local governments. The "*woonplaatsnamen*", that can be

literally translated as “names of places of residence”, appeared as the most interesting concept. The *woonplaatsen* are used in the everyday language, they are the toponyms people include when writing down an address. They are generally associated with population centres such as cities, town or villages and their surroundings.

13 To allow a data collection in a reasonable amount of time, it is very important to work on a limited number of entities. For this reason, we are focusing only on the top of the system of settlements. We must therefore distinguish place names that cover urban places from the rest, and do so for the entire period that is studied. Similarly to a previous cross-temporal analysis of the Dutch urban system (Van der Knaap, 1980), we decided to depart from the current situation and keep the list of units of analysis consistent throughout the period covered by the data collection. The main advantage of this choice is that the same basis is used for the entire period, meaning that the number of units of analysis does not change with time. Nonetheless, we acknowledge that there are also some drawbacks. A place that could not be considered as urban in the beginning of the period but only at the end will be included in the data for every period. In return, a place with a population dipping under the threshold during the period will not be included at all.

14 We created a database on population per *woonplaatsen* with census data available at postcode level for the year 2011[1] and the geocoding API from PDOK[2] to identify to which *woonplaats* the postcode was attached to. We kept only the *woonplaatsen* with more than 10,000 inhabitants. This threshold is often used by statistical agencies and scholars as the lower limit to define urban centres, and significantly reduces the number of places to query for. The result of this selection is a set of 317 Cities. Figure 2 presents their locations obtained from the Geonames database[3].

Figure 2: Location of the 317 cities for which data is collected

Classification of issues in place name recognition

15 City names, and more generally speaking place names, are subject to many ambiguities. These ambiguities can lead to important under- and overestimations when doing simple counts based on word frequencies. In a previous study (Meijers, Peris, 2018), different problems were identified in the case of the Dutch *woonplaatsen*. The most important sources of errors leading to false positives are listed below.

- Multiple meaning: some place names are similar to common names (i.e. “Huizen” = houses or “Dieren” = animals), verbs (i.e. “Leiden” = to lead, “Kampen”, to fight), and adjectives (i.e. “Houten” = wooden). This is the highest source of false positives.
- Homonymy: several places can have similar names. This is the case for “Katwijk”, which is at the same time a medium-sized coastal city in South Holland and a very small village in North Brabant. This can also be the case when a region and its most important city have the same name such as for Groningen and Utrecht.
- Family names: in quite some cultures, it is common to have a family name that relates to a place. In the list of cities that we are using for the data collection,

(Van/Van der) “Beek”, (Van) “Dongen” and (Van) “Doorn” are among the top 100 most frequent family names in the Netherlands[4].

- Organisations: place names are sometimes used by organisations, firms or institutions in their name. Our university, Delft University of Technology, is a case in point. Given the intimate connection of the most of these organisations with their place, one could argue that this is less a problem as news items using these organisation names will often be referring to something related to or happening in that place.

16 In terms of false negatives, there is one important source of errors:

- Multiple names: this can be the case when the old name coexists with the newer one (i.e. Den Haag/’s-Gravenhage and Den Bosch/’s-Hertogenbosch), when working on a multilingual corpus, or when places are also referred to with an abbreviation. The case of Alphen aan den Rijn, sometimes also spelled Alphen a/d Rijn, or simply referred to as ‘Alphen’ by some, can be mentioned.

A trade-off between computation time and precision level

17 Most of the issues mentioned above can be avoided by using a combination of Named-Entity-Recognition (NER) and disambiguation algorithms. NER is a subtask of Natural Language Processing (NLP) that aims to locate and classify entities from a given text into pre-defined categories. Named entities can be locations, persons, organisations, dates, measures (money, weight, distance, percent...), etc. In our case, considering the size of the dataset, such a method could not be used for every city as it would have taken months to perform NER on the entire corpus. We decided to go for a mixed technique to retrieve the data on cities in a reasonable amount of time. The issues that could occur with the list of 317 cities were listed and dedicated solutions were selected to handle these issues (Table 2). NER was used only for ambiguous cases. Different types of NER algorithms exist. Their efficiency largely depends on the types of entities that are targeted, the domain and the language. Studies applied to historical newspapers have shown that the level of performance of these algorithms can differ significantly (Ehrmann, Colavizza, Rochat, Kaplan, 2016; Mosallam, Abi-Haidar, Ganascia, 2014). In this project we have used the multiNER software[5], a NER set-up developed by the research department of the National Library of the Netherlands for the enrichment of several Dutch text corpora. This software is combining the outputs of 3 different NER packages in order to increase the accuracy of the recognition by using a certainty score. The leading package is the Stanford NER[6], which was previously trained manually using annotated sheets of Dutch historical newspapers (Neudecker, Wilms, Faber, van Veen, 2014). The two following packages are spaCy[7] and polyglot[8], both using a pre-trained Dutch NER-model.

Table 2: Issues in city name recognition and their solution.

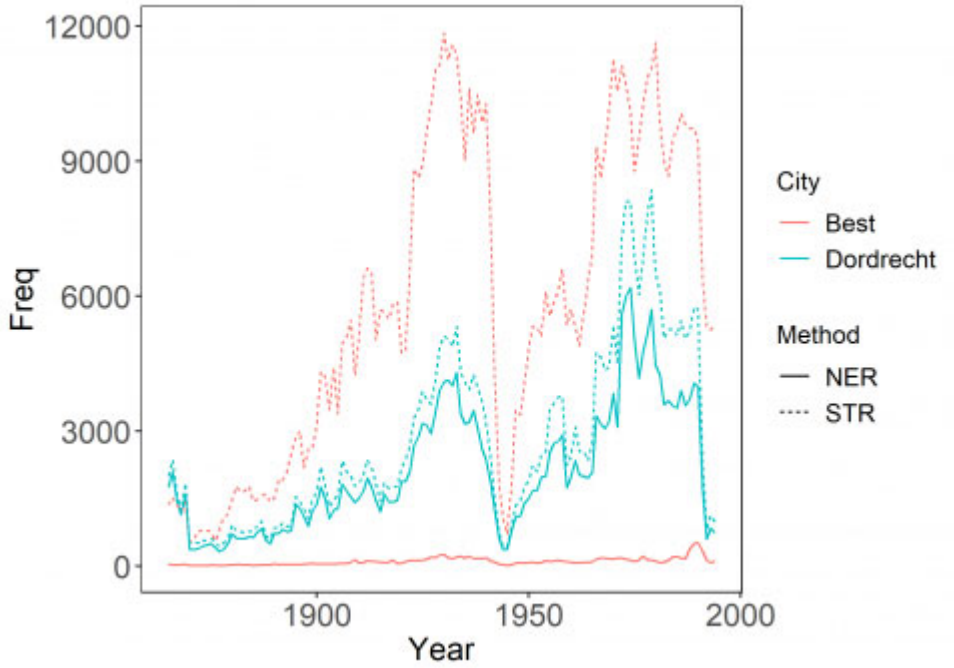
Problem	Frequency	Share (%)	Method	Implementation
Multiple meanings	23	7.3	NER	Yes
Homonymy	15	4.7	NER + *disambiguation*	No

Family names	3	0.9	NER	Yes
Organisation	*Unknown*	*Unknown*	NER	No
Multiple names	7	2.2	Multiple string queries	Yes
Unambiguous place names	274	86.4	String queries	Yes

18 Table 2 shows that the vast majority of city names is not ambiguous (86.4%) and does not require the use of NLP techniques. For these 274 cities, we performed SRU[9] queries using city names as simple search terms to retrieve the relevant articles from the corpus. This operation could be done in a reasonable amount of time.

19 In order to have an idea of the problem of false positives in the case of city names with multiple meanings, we measured the number of hits for two city names – one unambiguous and one ambiguous – in 21 different newspapers from the corpus. The two cities that were selected are Best, a small town close to Eindhoven which has a name that is a very common word in Dutch (the superlative of "better", like in English), and Dordrecht, a bigger city in South-Holland which has a very low chance of having false positives. Figure 3 shows that the case of Best manifests a considerable difference between the two techniques and require the use of NER. For this reason, in the case of the 23 cities with multiple meanings, we first collected the relevant articles via string queries using the SRU protocol and used the multiNER software to see whether the city was considered as a named entities. We kept the articles when 2 out of the 3 NER packages agreed that it was a named entity (a place name) indeed.

Figure 3: Comparison of two retrieving techniques for Best and Dordrecht.



Dashed lines represent the simple string queries (STR) performed via the SRU protocol, the solid lines shows the frequency after using the NER algorithm.

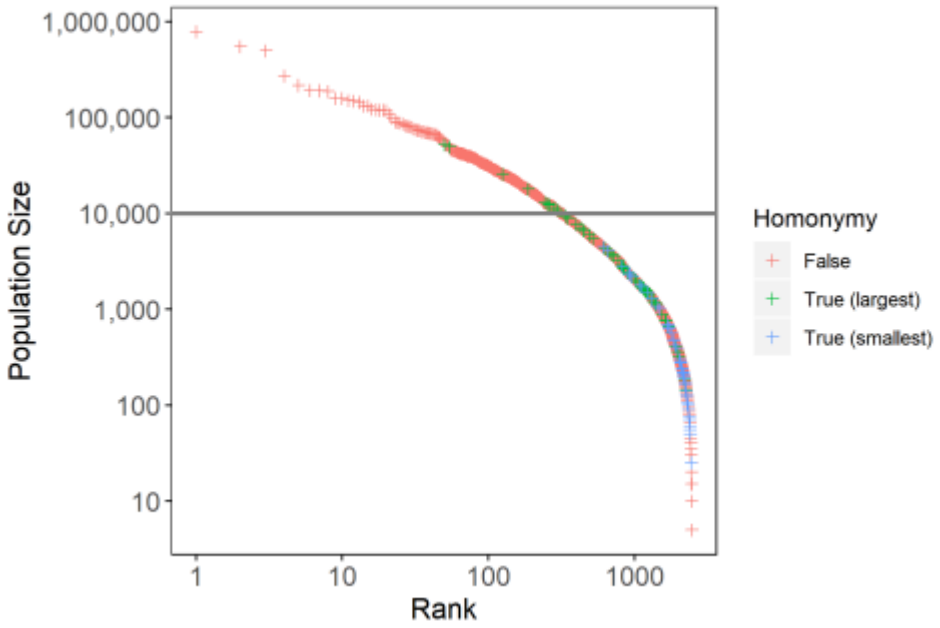
20 NER was also used for the 3 cities that occur often in family names. As we noticed some misclassification on the kind of named entities by the multiNER software, we kept only the articles with named entities that exactly matches the city name. This way, we could drop the names of people that are composed of a first name (or initials), a family name, and sometimes a prefix in between ("van", "de", "van der", etc.).

21 In the case of organisations, we could not apply NER because we had an insufficient knowledge of the organisations using the city names from the list. Such organisations would have been very difficult to identify considering the cross-temporal dimension. Moreover, as most of the time the use of a toponyms in the name of an organisation reflect relation with the place, this problem is not as important as multiple meanings and family names.

22 For cities with multiple names, multiple string queries via the SRU protocol were done. For example, we searched both for “Den Haag” and “‘s-Gravenhage” and aggregated the results afterward.

23 Finally, homonymy was the most difficult issue to handle. For a maximum level of precision, it would have been necessary to develop a specific disambiguation algorithm that uses the sentence around the named entity, the metadata of the newspaper (i.e. the place where the news item is published), as well as the importance of the possible places. However, we did not apply any disambiguation algorithm as the 15 cities from the list have homonyms of much smaller size (Figure 4). This limits the numbers of errors in their case.

Figure 4: Homonymy in settlement names.



The horizontal grey line represents the threshold above which the data is collected.

Data collection and structuration

24 The different steps of the data collection are summarized in Figure 5. They resulted in two files: one with the results of the data collection for the unambiguous city names (freq_count_STR.csv) and one for the ambiguous city names (freq_count_NER.csv). The file for unambiguous place names is structured the following way:

Table 3: Structure of the freq_count_str.csv file

ppn	city	type	year	freq
37631091X	Amsterdam	artikel	1871	347
37631091X	Amsterdam	advertentie	1871	149

25 The column *ppn* corresponds to a unique identifier given to each newspaper title. The column *city* is a character string corresponding to the city that is mentioned. The next variable, *year*, indicates the date. After that, *type* describes whether the city is mentioned in an article, an advertisement, some family announcements, or in the caption of an illustration. Finally, *freq* indicates the number of times this combination occurred. In this example, we can see that, according to the first line of the table, Amsterdam was mentioned in 347 articles of *De Maasbode*, a Rotterdam newspaper, in 1871.

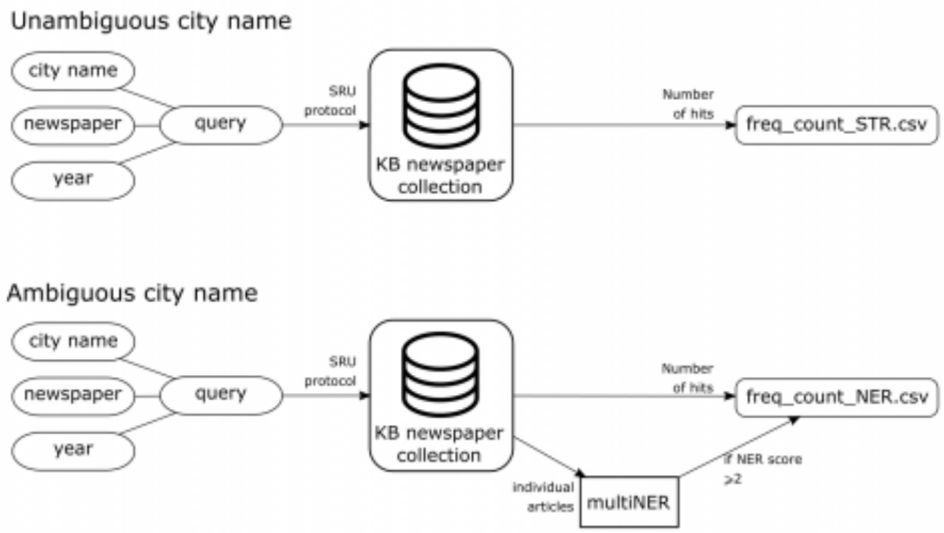
26 The file for ambiguous place names is structured almost the same way. The only difference is that additional to the frequency returned by the simple string query, there is an extra column with the number of hits after performing NER on the individual articles returned after the first query:

Table 4: Structure of the freq_count_ner.csv file

ppn	city	type	year	Freq_str	Freq_ner
400337266	*Leiden*	artikel	1914	89	21
400337266	*Leiden*	advertentie	1914	47	35

27 Additionally to these two files, the dataset contains also elements allowing to spatialise the data such as a file containing metadata on the newspapers, including the coordinates of the place where they were published (np_metadata.csv) and a file with the coordinates of the cities mentioned (cities_information.csv). Other files are also included such as freq_count_corps.csv, that contains the total number of items published in each year for every newspapers, which allows for example to standardise the data. More detailed descriptions of the files can be found in the metadata of the dataset.

Figure 5: Data collection algorithms



Technical Validation

28 In order to validate the accuracy of the dataset, we decided to manually assess the results of our algorithm on three different samples of news items containing in each

cases articles, advertisements and family announcements. The division in three sets was done on a temporal basis. We divided the period for which we have newspapers in three equal time slots (1869-1910, 1911-1952, 1953-1994) and randomly selected 50 news items between each of these time marks. This separation in different sets was done because we were aware that the quality of prints significantly improved during this period, affecting the efficiency of the automatic recognition of characters (OCR) used during the digitalisation of the newspapers. This selection resulted in three tables similar with the structure shown in Table 3. Each line corresponds to a news item. The first column contains an identification number, the second column the year of publication, the third column is the plain text of the digitalized item. The next three columns corresponds to the different steps of the place name identification. STR is the result of a simple string query for unambiguous place names, NER column is the result of a string query for the places that are in the list of ambiguous place names, and NER result is the outcome of the NER algorithm on the ambiguous place name. For example, in the case of the third row of Table 3, the string 'Goes' has been identified in the text of the news item, but the multiNER did not classify it as a place name, so it does not appear in the 'NER result column'.

Table 4: Structure of the sets used for sensitivity analysis

ID	YEAR	Plain text	STR	NER	NER result
1	1884	Keukenmeid of Huishoudster,Er biedt zich aan tegen half September een net Meïsja in eeu kleiu gezin als Of 31 1 ook is zij niet ongenegen eene ziekelijke Dame op te passen. Brieven franco, left. B, bij den Boekh. 1). RRAAIJ ENBRTNK, te Woerden. (5996)	Woerden		
2	1885	ZEE-MILITIE.De Burgemeeiter en Wethouders van Venloo nootfigen bij deze de lotelineen uit, die bij de Zee-Militie verlangen te dienen, zich daartoe bij hen aantemelden, ter plaatselijke Secretarie vóór den 1 April aanstaande. Venloo , den 12 Maart 1S86.			
3	1892	Burgerlgke Stand. GEHUWD: A. v. Dorp, jm. 31 en E. v. Vollenho ven, jd. 23 j., Pelikaanstraat 1. H. Pootman van Oije, wedr. 57 en M. L. Hazemijjer, wed. v. C. v. Hoek, 46 j., Hoogstraat 261. G.Kapsenberg, jm. 33 en J.A.v.der Goes,jd. 33}? L. Warande 106. J. H. Reidt, jm. 29 en A. F. v. Rjjn, jd. 26 j., Diergaardesingel 78. Huwelijks-Brieven en Verlovings-Circulaires worden gedrukt en spoedig afgeleverd, desverlangend geadresseerd ter drukkerij van het Nie uw sblad Goedkoop. Fijner papier naar keuze van i den besteller.		Goes	

29 We then counted the number of true positives, true negatives, false positives and false negatives to derive precision and recall indices for our three periods of time. These two indices are used to evaluate the outcome of an automated classification. The Precision *P* corresponds to share of relevant instances among the retrieved instances, and can be defined as:

$$P = \frac{tp}{tp + fp}$$

30 Where *tp* corresponds to the true positives and *fp* to the false positives. We also computed the recall *R*, which corresponds to the share of relevant instances that were

correctly retrieved. This index takes the following form:

$$R = \frac{tp}{tp + fn}$$

31 Where *fn* corresponds to the number of false negatives. Table 4 shows the results of these two calculations for the three different periods.

Table 5: Results of the precision and recall tests

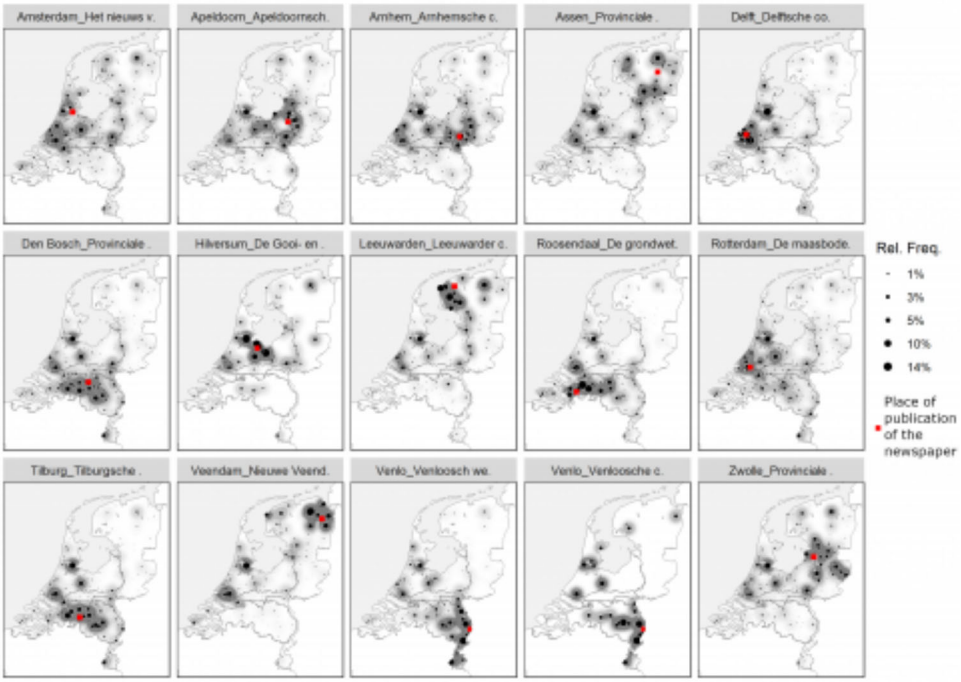
Period	*P*	*R*
1869-1910	0,91	0,92
1911-1952	0,98	0,96
1953-1994	0,99	0,88

32 The results of this validation process show an overall very good accuracy of our algorithm in the identification of place names in raw data. Most of the errors that we found in the randomly selected sample of articles were false negatives related to the quality of the OCR. This type of errors are almost unavoidable in quantitative analysis of digitalized texts. However, many efforts are being made by to constantly improve OCR quality.

Application: The information field of 15 Dutch cities in 1871

33 We present the maps resulting from the mapping of information field for 15 different cities in 1871 (Figure 5). Pred (1971) defines information fields as the total array of non-local contacts of individual places. Usually, these non-local contacts are likely to be high with nearby places, with which the frequency of interaction is important. In contrast, normally less information will likely be received from distant places. This pattern partly relates to the selection of information as being relevant for a group, given their pattern of interactions. Because of the considerable variability in the number of news items published in each newspaper we decided to plot the relative frequency of place-name mentions in comparison to the total number of news items published. To highlight the regional patterns in news coverage, we computed the Stewart potential with R package *SpatialPosition* (with an exponential function, span = 10000, beta = 3). These maps confirm the importance of distance for information flows as most of the attention is concentrated on the close-by cities and towns in 1871, with some attention to the big cities of the provinces of North and South-Holland. A more extensive study on the diffusion of information between the Dutch cities and its evolution over time can be found in Peris et al. (submitted).

Figure 6: Information field extracted from 15 local newspapers



Conclusion

34 Massive archives of digitalized text are full of geographic information. This is especially the case for archives of newspapers as these recorded the pulse of past societies. Quantitative analyses do not replace in depth readings, but they are a new way of looking at these sources and can reveal hidden patterns that appear only at the macroscopic scale. The reconstruction of the geography of information flows at different periods in time is such an hidden pattern that can only be observed through distant reading techniques. It allows to gain knowledge on the spatial organisation of territories through time.

35 However, extracting such patterns remains an important challenge from a methodological point of view. In this paper, we show the different steps that were needed to build a database on information flows between cities for a period of 125 years which we call DIGGER. It necessitated three main steps. The first one was to filter the different periodicals in order to keep only important ones. The second one was to select a sample of places that are consistent in terms of scale, toponomy and definition. Finally, the last step consisted in the building of an algorithm extracting accurate information from a massive dataset in a reasonable amount of time by using simple string queries and Named entity recognition (NER).

36 DIGGER can serve a wide variety of purposes. It has great potential for urban scholars to answer questions related to the dynamics of Dutch cities and the spatial diffusion of information, as well as by historians or media scientists interested in the geographical bias of news coverage. More generally, the methodology proposed in this data paper is of interest for people working on extracting geographic information from unstructured text data.

Dataset description

Language

37 English

Spatial coverage

38 317 cities and towns in the Netherlands (min y = 50.85, min x = 3.57, max y = 53.33, max x = 7.03). Reference system: ESPG 4326.

Temporal coverage

39 01/01/1869 – 31/12/1994

Format name and version

File	Description
cities_information.csv	geographical information on cities for which data has been collected
freq_count_corpus.txt	number of news items published by newspapers (per year/type)
freq_count_NER.csv	news flows between cities with ambiguous names and publication places of the newspapers (per year/type)
freq_count_STR.csv	news flows between cities with unambiguous names and publication places of the newspapers (per year/type)
np_metadata.csv	main information on the different periodicals included in the study
ppn_correspondance.txt	table to aggregate unique newspapers with multiple ppn code (unique identifier in Delpher database)
thesaurus_cities.txt	table to aggregate unique cities that have different names (ex. 's-Gravenhage = Den Haag)

Creation date

40 September 2018

Dataset creators

41 Peris A. (Department of Urbanism, Delft University of Technology, Delft, The Netherlands)
42 Faber W. J. (Koninklijke Bibliotheek, The Hague, The Netherlands)

Repository location

43 https://data.4tu.nl/repository/uuid:a14a1607-dafe-4a8a-aebc-d1c5cd66a588

Source

45 The digital archive of newspapers is accessible on the Delpher website (https://www.delpher.nl/). The data was created by querying the catalogue of the Koninklijke Biliotheek via a SRU protocol (http://jsru.kb.nl/sru/sru?query=).

Acknowledgements

46 This work was funded through a VIDI grant (452-14-004) provided by the Netherlands Organisation for Scientific Research (NWO), and through the researcher-in-residence program of the Koninklijke Bibliotheek, the national library of the Netherlands.

Bibliographie

Ehrmann M., Colavizza G., Rochat Y., Kaplan F., 2016, "Diachronic Evaluation of NER Systems on Old Newspapers",11.

Goodchild M., Li L., 2011, "Formalizing space and place", *CIST2011 - Fonder les sciences du territoire, Nov 2011, Paris, France. Proceedings du 1er colloque international du CIST,*, 177-183.

Grasland C., 2019, "International news flow theory revisited through a space–time interaction model: Application to a sample of 320,000 international news stories published through RSS flows by 31 daily newspapers in 2015", *International Communication Gazette*, 1748048518825091.

Knaap G. A. van der., 1980, *Population growth and urban systems development: a case study*. M. Nijhoff, 256 p .

Lansdall-Welfare T., Sudhahar S., Thompson J., Lewis J., Team F. N., Cristianini N., 2017, "Content analysis of 150 years of British periodicals", *Proceedings of the National Academy of Sciences*, Vol.114, No.4, E457-E465.

Maisonobe M., Eckert D., Grossetti M., Jégou L., Milard B., 2016, "The world network of scientific collaborations between cities: domestic or international dynamics?", *Journal of Informetrics*, Vol.10, No.4, 1025-1036.

Meijers E., Peris A., 2018, "Using toponym co-occurrences to measure relationships between places: review, application and evaluation", *International Journal of Urban Sciences*, 1-23.

Michel J.-B., Shen Y. K., Aiden A. P., Veres A., Gray M. K., Team T. G. B., et al., 2011, "Quantitative Analysis of Culture Using Millions of Digitized Books", *Science*, Vol.331, No.6014, 176-182.

Morse-Gagné E. E., 2011, "Culturomics: Statistical Traps Muddy the Data", *Science*, Vol.332, No.6025, 3535.

Mosallam Y., Abi-Haidar A., Ganascia J.-G., 2014, "Unsupervised Named Entity Recognition and Disambiguation: An Application to Old French Journals", 12-23in: P. Perner (Éd.), *Advances in Data Mining. Applications and Theoretical Aspects*. Lecture Notes in Computer Science. Cham, Springer International Publishing.

Neudecker C., Wilms L., Faber W. J., van Veen T., 2014, "Large-scale refinement of digital historic newspapers with named entity recognition", 16.

Peris A., Meijers E. J., van Ham M., 2019, "Information diffusion between Dutch cities: revisiting Zipf and Pred using a computational social science approach", *Submitted*.

Pred A. R., 1971, "Large-City Interdependence and the Preelectronic Diffusion of Innovations in the U.S.", *Geographical Analysis*, Vol.3, No.2, 165-181.

Schwartz T., 2011, "Culturomics: Periodicals Gauge Culture's Pulse", *Science*, Vol.332, No.6025, 35-36.

Tranos E., Kefalas P., 2018, "Digging into digital archives: the evolution of the digital economy in the UK", *Conference Paper*.

Van Kranenburg H. L., Palm F. C., Pfann G. A., 1998, "The life cycle of daily newspapers in the Netherlands: 1848–1997", *De Economist*, Vol.146, No.3, 475-494.

Zipf G. K., 1946, "Some Determinants of the Circulation of Information", *The American Journal of Psychology*, Vol.59, No.3, 40-421.

Notes

1 http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=81310ned&D1=0&D2=a&HDR=T&STB=G1&VW=T

2 https://github.com/PDOK/locatieserver/wiki/API-Locatieserver

3 https://www.geonames.org/

4 http://www.cbgfamilienamen.nl/nfb/documenten/top100.pdf

5 https://github.com/KBNLresearch/multiNER

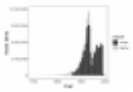
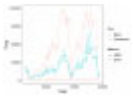
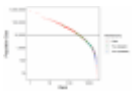
6 https://nlp.stanford.edu/software/CRF-NER.shtml

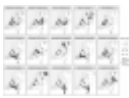
7 https://spacy.io/

8 http://polyglot.readthedocs.io/en/latest/

9 http://www.loc.gov/standards/sru/

Table des illustrations

	**Titre**	Figure 1: News items per year in Delpher and in the sub-corpus
	URL	http://journals.openedition.org/cybergeo/docannexe/image/33747/img-1.png
	Fichier	image/png, 25k
	**Titre**	Figure 2: Location of the 317 cities for which data is collected
	URL	http://journals.openedition.org/cybergeo/docannexe/image/33747/img-2.png
	Fichier	image/png, 156k
	Titre	Figure 3: Comparison of two retrieving techniques for Best and Dordrecht.
	**Légende**	Dashed lines represent the simple string queries (STR) performed via the SRU protocol, the solid lines shows the frequency after using the NER algorithm.
	URL	http://journals.openedition.org/cybergeo/docannexe/image/33747/img-3.png
	Fichier	image/png, 41k
	Titre	Figure 4: Homonymy in settlement names.
	**Légende**	The horizontal grey line represents the threshold above which the data is collected.
	URL	http://journals.openedition.org/cybergeo/docannexe/image/33747/img-4.png
	Fichier	image/png, 25k
	Titre	Figure 5: Data collection algorithms
	**URL**	http://journals.openedition.org/cybergeo/docannexe/image/33747/img-5.png
	Fichier	image/png, 198k

$P = \frac{tp}{tp + fp}$	**URL**	http://journals.openedition.org/cybergeo/docannexe/image/33747/img-6.png
	Fichier	image/png, 1,2k
$R = \frac{tp}{tp + fn}$	**URL**	http://journals.openedition.org/cybergeo/docannexe/image/33747/img-7.png
	Fichier	image/png, 1,3k
	**Titre**	Figure 6: Information field extracted from 15 local newspapers
	URL	http://journals.openedition.org/cybergeo/docannexe/image/33747/img-8.png
	Fichier	image/png, 532k

Pour citer cet article

Référence électronique
Antoine Peris, Willem Jan Faber, Evert Meijers et Maarten van Ham, « One century of information diffusion in the Netherlands derived from a massive digital archive of historical newspapers: the DIGGER dataset », *Cybergeo : European Journal of Geography* [En ligne], Data papers, document 928, mis en ligne le 14 janvier 2020, consulté le 14 janvier 2020. URL : http://journals.openedition.org/cybergeo/33747

Auteurs

Antoine Peris
Department of Urbanism, Delft University of Technology, Delft, Netherlands
a.f.t.peris@tudelft.nl

Willem Jan Faber
Koninklijke Bibliotheek, National Library of the Netherlands, The Hague, Netherlands
willemjan.faber@kb.nl

Evert Meijers
Department of Urbanism, Delft University of Technology, Delft, Netherlands
e.j.meijers@tudelft.nl

Maarten van Ham
Department of Urbanism, Delft University of Technology, Delft, Netherlands
School of Geography and Sustainable Development, University of St Andrews; St Andrews, Scotland
m.vanham@tudelft.nl

Droits d'auteur